ТАРТУСКИЙ УНИВЕРСИТЕТ

АКТУАЛЬНЫЕ ПРОБЛЕМЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

ТАРТУ 1990

ТАРТУСКИЙ УНИВЕРСИТЕТ

АКТУАЛЬНЫЕ ПРОБЛЕМЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Тезисы докладов Всесоюзной конференции
в гор. Тарту 29-31 мая 1990 г.

Тарту 1990

Организаторы конференции:

Тартуский университет
Московский государственный университет
Институт языкознания АН СССР
Институт языка и литературы АН Латвии

Оргкомитет:

Ю.А. Тулдава (председатель); К.Я. Лепа, А.А. Поликарпов (сопредседатели); С.В. Райтар (секретарь); Ю.А. Викс, В.А. Дризуле, Р.Г. Котов, Ю.Н. Марчук, Х.Я. Ыйм

АЛАТОРЦЕВА С.И., БУЦЕВА Т.Н.

ПРОБЛЕМЫ РАЗРАБОТКИ АВТОМАТИЗИРОВАННОГО СЛОВАРЯ РУССКИХ НЕОЛОГИЗМОВ

I. Теоретически структура МФРЯ уже в основном определилась. В процессе создания конкретных подфондов необходимо больше внимания уделять всему комплексу лингвистических проблем, которые тот или иной структурный элемент МФРЯ может решать.

2. По замыслу словарно-грамматический подфонд МФРЯ должен включить в себя все академические толковые словари. Среди них можно выделить блок словарей, отражающих лексический состав современного русского языка. Они же являются базой для лексикографического описания лексики последующих периодов. Неологические словари - неотъемлемая структурная часть предполагаемого подфонда.

3. Сложилось три типа лексикографических изданий русской неологии, различия которых необходимо учитывать при создании АСРН - автоматизированного словаря русских неологизмов. Все словари неологизмов по отношению к существующим словарям - дифференциальные, дополняющие их. Они не только регистрируют и описывают новообразования определенного периода, но и обобщают результаты процессов узуализации этой лексики.

4. Автоматизированный словарь может строиться как:
а) словник с указанием лексикографического источника и параметрических характеристик, извлеченных из словарных статей,
б) полный текст словаря, введенный по определенным полям с целью извлечения конкретного материала для редактирования, лексикологических исследований и т.п.

5. Целевое назначение АСРН - служить основой для формирования словников словарей разного типа, для создания дифференцированных словарей неологизмов, для лексикологических исследований и т.п.

6. В Словарном отделе ЛО ИЯ АН СССР сделан пробный фрагмент АСРН по типу а и начата работа над АСРН по типу б по программе, используемой там же для создания автоматизированного МАС.

АНДРЕЕВ С.Н., КОВАЛЬКОВА И.В.

КОМПЬЮТЕРНЫЙ АНАЛИЗ ПРИЗНАКОВОЙ СИСТЕМЫ АНГЛИЙСКИХ ПРОИЗВОДНЫХ ГЛАГОЛОВ

Одним из основных направлений анализа структуры объекта является построение классификации его элементов. Проведение классификационного анализа позволяет создать конструкт, моделирующий совокупность эмпирических отношений.

Группировка признаков языка, как правило, осуществляется при отсутствии достаточной информации для того, чтобы сформулировать априорно архетип классов и составить их полный список, т.е. дать классификационную схему. Систематическая классификация элементов языка поэтому в большинстве случаев требует дедуктивного многомерного подхода, изменения признакового пространства в ходе самого исследования и установления степени устойчивости получаемых результатов.

Исследование такого вида оказывается возможным при условии создания базы данных, которая обладала бы следующими чертами: (I) большой объем информации (обрабатываемой и хранимой); (2) возможность оперативной селекции исходного материала по разнообразным признакам и их комбинациям и значительная статистическая обработка исходных и промежуточных данных. В докладе приводятся результаты многомерного анализа (факторного и кластерного) соотношений разноуровневых признаков английских производных глаголов. Программа анализа материала включает основную часть, обеспечивающую построение массива данных, а также две подпрограммы, позволяющие осуществлять оперативный ввод признаков, их статистическую обработку, вывод и хранение полученных результатов. Нами использовался персональный компьютер, типа IBM PC/AT. Всего в базу данных было введено более 2600 производных аффиксальных глаголов и 2400 аффиксальных глаголов, обладающих условной и дефектной членимостью. К анализу на различных этапах привлекалось около I00 разноуровневых признаков. В качестве меры связи признаков использовались коэффициенты корреляции Пирсона-Бравэ, Коула, меры расстояния и др. В докладе приводятся основные сопоставительные данные, отражающие степень устойчивости группировки признаков в зависимости от типа близости исходных таблиц и количества используемых признаков.

АНДРЕЕВСКАЯ А.В.

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ НА БАЗЕ МАШИННОГО СЛОВАРЯ КОРНЕВЫХ СЛОВ (К ВОПРОСУ О СВЯЗИ КОЛИЧЕСТВЕННЫХ ХАРАКТЕРИСТИК ПОЛИСЕМИИ И ПРОДУКТИВНОСТИ).

1. Данные, анализируемые ниже, содержатся в машинном Словаре корневых слов русского языка XI-XX веков (СКС), созданном автором совместно с О. Г. Климовой в рамках работ над Машинным Фондом русского языка в Институте русского языка АН СССР Объем основных баз словаря - 7 МБайт на жестком магнитном диске ПЭВМ совместимой с **IBM PC** в базе данных типа **FOXBase**. Словарь включает в себя 5 858 словарных статей, каждая из которых состоит из семи зон, объединяющих в общей сложности 20 параметров. Эти параметры характеризуют графический облик слова, наличие у него фонетических и грамматических вариантов, грамматику слова, число его значений и их отнесенность к тому или иному семантическому полю, происхождение, продуктивность и употребительность корневого слова, а также некоторые сведения об истории его корня. В каждую зону эксплицитно или по умолчанию входят хронологические параметры, которые отражают время первой и последней Фиксаций как самого слова, так и всех его признаков. Путем обработки всего массива корневых слов по хронологическим параметрам были получены вспомогательные базы **STRAT** 1,...,5, позволяющие оперативно анализировать лексику определенной эпохи; аналогичные базы созданы и для значений (**MSTRAT** 1,...,5).

2. Корреляционный анализ корневых слов по данным СКС проводится на основе матриц для выбранной пары признаков, которые составляются по результатам сплошной машинной обработки множества корневых слов определенной эпохи. Дальнейший анализ данных ведется средствами стандартных пакетов статистической обработки. Например, таким образом были исследованы количественные характеристики полисемии и продуктивности корневых слов.

Было установлено, что между числом производных и числом значений существует несомненная корреляция на протяжении рассматриваемых девяти веков (XI-XX). Корреляционная зависимость близка к линейной. Наиболее явно связь между парой анализируемых параметров проявляется при небольших значениях обоих признаков: числе дериватов до 10 (максимальное значение для корневых слов - 25-30), количестве значений до 5-8 (соответ-

ствующие максимумы 13-20, в зависимости от периода). При более высоких значениях обоих признаков корреляция становится слабее. Это связано с тем, что на этом интервале фиксируется незначительное число объектов-корневых слов и их распределение, соответственно, более чувствительно к случайным влияниям. В целом коэффициент корреляции r(1) принимает следующие значения в зависимости от времени 1: r(X1-X1V вв.)=0.307, r(XV-XVII вв.)=0.278, r(XVIII в.)=0.301, r(X1X в.)=0.322, r(XX в.)=0.343. Наблюдаемые значения критерия Стьюдента для значимости этих корреляций растут с течением времени и существенно выше критических. Таким образом, гипотеза о наличии корреляции статистически значима.

В то же время, при наблюдаемом уровне коэффициента корреляции (около 0.3) нельзя говорить о жесткой детерминированности одного параметра другим. Можно предположить, что корреляция между числом значений и числом дериватов обусловлена не столько взаимным влиянием этих двух признаков, сколько воздействием на них какого-либо третьего, более значимого параметра (или параметров). Вероятно таким параметром является частота корневого слова. СКС позволяет проверить эту гипотезу для современного русского языка.

Известно, что существует пропорциональная зависимость между употребительностью слова и его многозначностью. С другой стороны, для корневых слов XX века установлено наличие сильной корреляции между числом дериватов и частотой (коэффициент корреляции r=0.581). Таким образом, зависимость количественных характеристик продуктивности и многозначности друг от друга значительно слабее, чем каждой из них в отдельности от частоты. Из этого можно заключить, что употребительность является одним из тех признаков, которые обуславливают силу связи между количеством дериватов и числом значений корневого слова. Возможно, что это не единственный фактор, воздействующий на пару исследуемых параметров. (В частности, можно ожидать корреляций между квантитативными характеристиками полисемии и продуктивности и длиной слова или его возрастом). Однако, без сомнения, частота является наиболее значимым из рассматриваемых признаков. Дальнейший анализ материалов Словаря корневых слов русского языка X1-XX веков описанным выше методом машинной обработки позволит строго ранжировать указанные факторы.

АНИКИНА Н.В.

БАЗА ДАННЫХ "СЕМАНТИКА ПРЕДЛОГОВ"

Рассматривается организация и применение машинной базы данных, содержащей сведения о семантике русских и английских предлогов. База данных состоит из 2-х частей - словарной и текстовой.

I-ая часть базы данных - это англо-русский автоматический словарь предлогов. В нем в словарной статье каждого предлога приводится перечень его семантических признаков. Кодирование этих признаков осуществлено на основе разработанной нами иерархической системы представления семантики предлогов, включающей II4 семантических признаков.

2-ая часть базы данных - это фрейм, содержащий около 500 минимальных семантических контекстов, в каждом из которых актуализуется некоторый конкретный семантический признак предлога. Каждый такой контекст является семантическим каркасом, по которому строится достаточно большое количество реальных текстовых предложных словосочетаний.

Рассматриваемая база данных может использоваться для представления семантики предлогов не только в английском и русском, но и в других языках.

На ее основе могут выполняться следующие действия:

- выбор всех семантических признаков предлога;
- определение семантического признака, актуального в конкретном предложном словосочетании;
- определение всех предлогов некоторой семантической группы и предлогов-антонимов;
- определение всех семантических контекстов, в которых реализуется заданный признак предлога, и синтез соответствующих им словосочетаний-примеров, и др.

Основные области применения рассматриваемой базы данных:

- машинный перевод (приводятся результаты англо- и испано-русского машинного перевода предлогов);
- машинные фонды языков;
- компьютеризация обучения языкам.

АНИСИМОВ А.И.

ИСПОЛЬЗОВАНИЕ ЯЗЫКА ФОРТ В КОМПЬЮТЕРНОЙ ЛЕКСИКОГРАФИИ

Составители различных словарей в ходе долгой практики пришли к наиболее удобным и обычным для человека определениям слов (понятий). Можно предположить, что определения понятий-данных для пользователей лексикографической экспертной системы (ЛЭС) целесообразно построить по аналогии с определением понятий, принятым в толковом словаре, а определение понятий-операций - по аналогии с определениями новых слов в языке ФОРТ.Язык ФОРТ получил широкое распространение среди практиков-программистов и вызвал интерес теоретиков рядом новаторских черт.Популярность его связана со многими привлекательными свойствами ФОРТА, среди которых не последнее место занимает принципиальная простота аппаратной реализации его элементов, простота компактного представления программ в виде шитого кода /1/, простота его расширения путем добавления в словарь новой словарной статьи - - специальной структуры данных, состоящей из поля имени, поля связи, поля кода и поля параметров.

С другой стороны разработка программного обеспечения инструментальной системы, автоматизирующей процесс производства ЛЭС, базируется на общей модели словарной статьи, ориентированной на разбор по составляющим компонентам, на методике разметки словарной статьи, адекватной модели ее структуры, и на специально построенной формальной грамматике, описывающей структуру словарной статьи для филологических словарей различных типов /2/.

В сообщении предлагается использовать в качестве языка представления знаний ЛЭС расширение языка ФОРТ, реализованное в Тартуском университете /3/. Это расширение радикально отличается от Пролога своим сугубо императивным характером. Тем не менее облачаться в форму декларативного стиля ФОРТ также может. Предлагаемое дальнейшее расширение ФОРТА, названное ФОРТ++, позволяет в удобном для пользователя виде

представить знания так называемыми интегрированными словарными статьями (ИСС).

Основными компонентами ЛЭС словарного типа являются : интеллектуальный банк данных (ИБД), организованный в виде стека различных филологических словарей; правила, задающие знания о предметной области; система программных процессоров управления. Процесс работы ЛЭС состоит в применении правил к ИБД. Система управления выбирает, какие правила и в какой последовательности применять. Система прекращает работу, когда достигаются заданные цели. Задача заключается в описании базисных функций, в частности, базовых операций над словарями таким образом, чтобы их суперпозиция представляла собой программу на расширенном языке ФОРТ++, выполнение которой было бы обработкой исходного обращения к ИБД. Связи между понятиями и ИСС в своей совокупности образуют не иерархию, как обычно, а граф. Это дает возможность естественным образом организовать параллельные вычисления при использовании знаний, содержащихся в словарях ИБД ЛЭС.

Целесообразность выбора расширения ФОРТА в качестве языка для суперкомпьютера пятого поколения и ИСС в качестве основного элемента словарных баз ИБД подкрепляется тем, что уже имеются и активно используются сети мультитранспьютерных систем и процессоров шитого кода, обладающие быстродействием, наращиваемой архитектурой и способностью непосредственно интерпретировать языки функционального программирования, в том числе, языки типа ФОРТ++ /4/.

ЛИТЕРАТУРА

I. Баранов С.Н.,Ноздрунов Н.Р.Язык Форт и его реализации.- Л.: Машиностроение. 1988.

2. Колодяжная Л.И.Автоматизированная лексикографическая система УНИЛЕКС. -М.:Изд-во Моск.ун-та, 1987.

3. Томбак М.О. Форт 83/32. Отчет. Номер гос.регистрации 01870090451.-Тарту: ТГУ,1989.

4. Голден Дж.,Мур Ч.,Брод Л. Быстродействующий однокристальный процессор,исполняющий команды,написанные на Форте.- Перевод № СР- 20035.- М.: ВЦП,1988.

АНИСИМОВ А.И., ГРАМОЛИН В.В., КОЛОДЯЖНАЯ Л.И.

ОБУЧАЮЩАЯ ЛЕКСИКОГРАФИЧЕСКАЯ ИГРА НА КОМПЬЮТЕРЕ

В начале было Слово...
Иоан. 1.1.

Сейчас все чаще компьютерные игры помимо развлекательной снабжаются социально-значимой целью и наполняются духовно-ценным содержанием. В сообщении впервые рассматривается новый класс портативных компьютерных игр, основной целью которых является обучение различным аспектам лексикографических знаний. В основу программных средств этого класса положены новые концепции построения обучающих игр, методики их развития, единый принцип составления формального описания словарной статьи любого филологического словаря и единый механизм сведения конкретных лексикографических операций к четырем базовым операциям. Приведены результаты проектирования и макетирования первой игры этого класса PUN, обучающей лексике русского и английского языков.

Игра спроектирована для реализации в портативном варианте: карманном - размером приблизительно 50x100x10. Своим внешним видом она напоминает известные электронные игры типа "Ну погоди". Игра разработана с учетом использования жидкокристаллического экрана, микропроцессора КБ1013ВК7-2 (или SM-5.11.7-2 для управления индикацией на нем), процессора шитого кода (типа NC6000 фирмы "НОВИКС" с 144 выводами на корпус, частотой 10 мгц) и 256 страниц по 64 килобайт для трех пространств памяти (48 Мб). Игра имеет четыре игровых режима (по два для англо-русского словаря и словаря-перевертыша), режимы обучения числительным и, попутно, таблице умножения, справочно-обучающий режим, а также режим часов и будильника.

В начале игры на экране появляется первое английское слово. В это время сверху и сбоку по экрану начинают "сыпаться" буквы кириллицы для набора слова русского перевода. Если никаких действий не предпринимать, то первая же буква, опустившись на место ответа, приведет к неверному результату. Для получения ответа, персонаж "ученика", управляемый игроком, отбивает ненужные буквы, нужные пропускает, чтобы в начале было слово, являющееся правильным переводом на русский язык.

При правильном ответе возникает изобразительная и звуковая сигнализация поощрения. Слова, на которые даны неправильные ответы, через некоторое время выставляются вновь. В процессе игры ведется счет: при правильном переводе увеличиваются премиальные очки, при неправильном - штрафные. Управление игрой осуществляется не более, чем четырьмя клавишами, которые манипулируют изображением "ученика" на экране. Слова выставляются попеременно на английском языке или беруться заглавные слова из статьи словаря-перевертыша, причем поступают они в псевдослучайном порядке, но в направлении от простого к сложному, от общеупотребимой лексике к конкордансу и симфонии, соответственно.

При таком подходе игрок, хотя и оперирует буквами, совсем не нуждается в буквенно-цифровой клавиатуре, что позволяет вести игру не отрывая взгляда от экрана. Это создает комфортную психологическую ситуацию, снижает стоимость игры, повышает ее надежность. Переход в другой игровой режим приводит к тому, что начинают "сыпаться" слова, начинающиеся с букв, которые были выбраны игроком в нормальном режиме. Помимо этих двух игровых режимов - нормального и ускоренного - имеются еще четыре: справочно-обучающий, режимы обучения таблице умножения, и режим часов. В справочно-обучающем режиме весь словарь на экране демонстрируется так: после установки заглавного слова отображаются почти все компоненты словарной статьи, кроме зоны перевода. Дается некоторая пауза, лишь после которой - ответ. За время паузы учащийся может увидеть все синонимы и тексты из зоны примеров словарной статьи, предположить ответ, а затем удостовериться как слово перевода записывается на самом деле.

Режим часов - это одновременная индикация как циферблата со стрелками, так и цифрового обозначения времени.

В режимах обучения таблице умножения: в нормальном - "сыпятся" не буквы, а цифры и числа, в ускоренном-числительные.

Для обучения более подготовленных учеников предполагается использовать не только переводные словари, но и филологические словари других типов, например, синонимов, толковые, другие и, конечно, конкорданс и симфонию.

Действующий макет процессора шитого кода уже построен из секционированных микропроцессоров. Важнейшие компоненты приведенного выше сценария обучающей игры FUN для режимов обучения таблице умножения также проверены на микропроцессорном макете.

АНИСИМОВ М. А.,
СУРКИС А. С.,
ЯБЛОНСКИЙ С. А.

СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР ПРОЦЕССОРА РУССКОГО ЯЗЫКА РУССИКОН-1

Существующие системы синтаксического анализа предложений русского языка [1] реализованы на ЕС ЭВМ, что не позволяет осуществить их прямой перенос на персональные ЭВМ. Кроме того, большинство алгоритмов синтаксического анализа требует создания дополнительных синтаксических зон в словарных статьях машинного словаря русского языка. При размере словаря 80 - 100 тыс. словообразовательных основ объем предварительных трудозатрат по созданию синтаксических зон становится чрезмерно большим. Поэтому в качестве базовых методов в синтаксическом анализаторе процессора русского языка РУССИКОН-1 были выбраны подходы [1,2]. Данные методы не требуют дополнительной обработки словаря и большого перебора вариантов при синтаксическом анализе предложений русского языка и при этом используют в основном лишь грамматическую информацию слов предложения, получаемую после выполнения морфологического анализа. С другой стороны указанные подходы обеспечивают сравнительно невысокое качество построения деревьев зависимости для сложных предложений.

Результатами работы синтаксического анализатора является:

- разрешение морфологической омонимии на этапе предсинтаксического анализа;
- выделение именных словосочетаний;
- получение деревьев зависимостей словосочетаний.

Это достигается путем выполнения следующих основных шагов:

1. Предсинтаксический анализ предложения.

На этом этапе осуществляется проверка более 200 правил согласования слов, к которым относятся:

- правила согласования предлогов и падежей стоящих за ними слов;

- правила согласования членов предложения;
- модели управления глаголов и пр.

Предсинтаксический анализ позволяет устранять до 80% случаев морфологической омонимии слов в предложении.

2. Синтаксический анализ предложений.

Для анализа простого предложения используется метод В. Г. Сухотина [2].

При построении оптимального дерева зависимостей предложения используется оценочная функция, определяющая связи слов в предложении на основе матрицы вероятностей связей между словами разных грамматических классов. Вероятности в матрице вычисляются для встречаемости грамматических классов в непосредственной близости и в пределах простого предложения. При построении связей между словами разных участков в пределах одного сложного предложения связи определяются путем выделения вершин локальных поддеревьев всех участков предложения, ограниченных знаками препинания, сочинительными союзами, глаголами и некоторыми другими словами. Связи между участками формируются с использованием алгоритмов, аналогичных [1]. Таким образом, методу [1] отводится роль построения локальных поддеревьев и выделения их вершин.

Синтаксический анализатор реализован на основе языка СИ и языка сетей переходов [3] в среде MS DOS.

Л И Т Е Р А Т У Р А

1. Итоги науки и техники / сер. Информатика, т.8 - М.: ВИНИТИ, 1984.

2. Сухотин Б. В. Оптимизационные методы исследования языка. - М.: Наука, 1976.

3. Суркис А. С., Яблонский С. А. Язык сетей переходов // Программное обеспечение новой информационной технологии. - Калинин, 1989.

АПОЛЛОНСКАЯ Т.А.

ОРГАНИЗАЦИЯ СЛОВАРНОЙ ИНФОРМАЦИИ ПРИ АВТОМАТИЧЕСКОЙ ПЕРЕРАБОТКЕ ТЕКСТА

Одной из причин сближения теоретического языкознания и теоретического аспекта прикладного языкознания является использование функционального подхода к анализу языковых явлений. Подтверждением этого сближения служит все возрастающий интерес к основному методологическому и теоретическому вопросу языкознания - проблеме взаимоотношения языка - мышления - речи. В рамках наметившегося сближения особого внимания заслуживает исследование особенностей взаимодействия грамматики и словаря как в теоретической, так и в прикладной лингвистике.

Задача настоящего доклада состоит в том, чтобы показать взаимодействие грамматики и словаря при решении задач автоматической переработки текста /АПТ/. С целью обеспечения их взаимодействия при разработке систем АПТ используется структурный подход, осуществляемый по нисходящей схеме "сверху вниз". Такой подход предполагает наличие открытой модульной системы, четкое представление уровней анализа, установление их иерархии, определение задач на каждом из уровней, а также знание объема и типов информации, помещаемой в словарную статью.

В соответствии с этим процедура получения перевода разбивается на уровни: предложения, функциональных компонентов, функциональных сегментов, лексического уровня. При этом каждый из предыдущих уровней становится фундаментом для последующего более продвинутого уровня.

При такой постановке задачи разработчик должен четко определить набор признаков, обеспечивающих анализ и синтез на каждом из выделенных уровней. Эти задачи определяют объем, структуру словарной статьи /СлСт/ лексической единицы. Взаимная корелляция грамматики и словаря на каждом из выделенных уровней демонстрируется на примере анализа СлСт существительного. СлСт существительного включает IO позиций, содержащих информацию о принадлежности к классу слов, к типу омонимии, к семантическому классу, о наличии формальных, морфологических, функциональных, валентностных и лексико-грамматических характеристик.

Интерес представляет описание механизма использования

словарной информации в ходе анализа и перевода ЛгвЕ.

На уровне лексического анализа осуществляется приписывание каждой входной единице текста всего набора словарной информации, содержащейся в словарях словоформ и оборотов. Здесь же осуществляется процедура распознавания аналитических конструкций. Результатом проработки лексического уровня является получение пары входной-выходной лексической единицы со всем набором словарных признаков.

На следующем уровне - уровне функциональных сегментов распознаются именные сегменты равные простой именной группе. Процедура анализа именной группы осуществляется с ориентацией на информацию о морфологических, лексико-синтаксических и лексико-грамматических характеристиках, содержащихся в СлСт. Результатом работы на уровне функциональных сегментов является получение пары входной-выходной функциональный сегмент, содержащий набор признаков, свойственных этому уровню. На следующих уровнях может происходить коррекция имеющихся признаков и приписывание новых признаков, свойственных структурам более высокого уровня.

На уровне функциональных компонентов происходит распознавание сложной именной группы. Эта процедура осуществляется на основе валентностных характеристик ядерного элемента группы. Выбор переводного эквивалента может зависеть не только от валентности имени, выступающего в качестве ядерного элемента, но и от принадлежности его к определенному семантическому классу. Таким образом, здесь используются семантико-синтаксические и лексико-грамматические характеристики. Результатом работы этого уровня является цепочка функциональных компонентов, имеющих определенный набор признаков, необходимых для анализа на следующем уровне - уровне предложения.

На уровне предложения функциональные компоненты получают статус членов предложения. Именно здесь происходит коррекция перевода именного функционального компонента и выбор переводной структуры целого предложения.

Таким образом, выбранный подход к разработке систем АПТ позволяет представить весь объем задач, стоящих перед исследователем, и создает условия для разработки систем, прогнозирующих анализ вновь возникающих ситуаций, на основе уже проработанных ситуаций.

БАСКЕВИЧ В.М.

АВТОМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СЕМАНТИКО-ПАРАДИГМАТИЧЕСКИХ ОТНОШЕНИЙ В ЛЕКСИКЕ

Применение в современном языкознании квантитативных методов исследования позволило "измерить" лексико-семантическую систему языка, объединить ее элементы в сложную, поддающуюся автоматическому анализу систему взаимоотношений. В результате получена многоуровневая модель, в которой в гармоничном соответствии друг с другом находятся синтагматические и парадигматические отношения.

Проведен статистический анализ синтагматических связей прилагательных с существительными (I6 тыс.примеров) в текстах газетно-публицистического стиля современного немецкого языка. Выделены семантические разряды прилагательных и существительных, выборочные частоты сочетаемости которых были зафиксированы в исходной таблице. С помощью коэффициента линейной корреляции (r) измерялась связь между разрядами прилагательных. Программа для ЭВМ была составлена таким образом, чтобы получить не только коэффициенты корреляции между разрядами прилагательных, но и определить тесноту и направление связи.

Анализ показал, что, например, прилагательные, обозначающие общественно-политические отношения, наиболее тесно коррелируют с прилагательными с семантикой пространственных характеристик. Это означает, что прилагательные общественно-политических отношений сочетаются с теми разрядами существительных, что и прилагательные, обозначающие пространственные характеристики: politische Arbeit и gesamte Arbeit ; internationale Welt и ganze Welt и т.д. Прилагательные интенсивности наиболее тесно коррелируют с прилагательными динамичности, прилагательные температуры - с прилагательными интенсивности.

Наиболее тесно коррелируют те разряды прилагательных, которые имеют наибольшую общность синтагматических связей с разрядами существительных. Следовательно, источником семантических корреляций является общность синтагматических связей.

Становится вполне очевидным, что синтагматические и парадигматические отношения существуют не разрозненно, а в комплексе, в единстве и взаимодействии, т.е. так, как объективно предопределено самой системой языка.

Парадигматический анализ позволил на основе количественных измерений семантико-корреляционных отношений описать устройство лексической подсистемы газетно-публицистического стиля как некую

подструктуру, получить модель с большим числом взаимосвязей и наглядно изобразить ее на рисунке (рис.I).

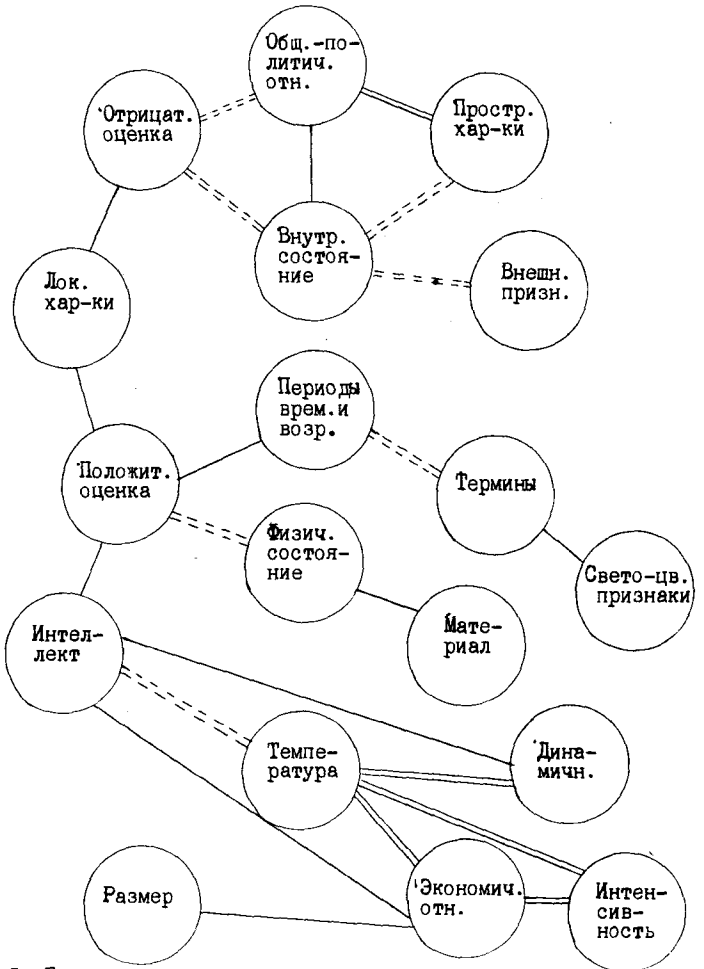


Рис. I. Граф корреляций между семантическими разрядами прилагательных Условные обозначения: ═══ сильная связь; ——умеренная связь; - - - - слабая связь.

БАХМУТОВА И.В., ГУСЕВ В.Д., ТИТКОВА Т.Н.

ВОЗМОЖНОСТИ АНАЛИЗА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ С ПОМОЩЬЮ АЛГОРИТМОВ ДЕШИФРОВОЧНОГО ТИПА

Авторы в течение ряда лет занимаются анализом слитных неструктурированных текстов (генетических, музыкальных, двоичных последовательностей) с целью выявления функционально значимых структурных единиц текста на различных иерархических уровнях. В ситуации, когда элементы структуры неизвестны, можно рекомендовать некоторые стандартные заготовки для них, основанные на понятии повтора. Состав, количество и расположение повторов в тексте несут важную информацию о его структуре.

В используемой нами системе описания текстов фиксируется полный спектр повторов, начиная с самых коротких (длины I) - до самых длинных. Повторы классифицируются на совершенные (идентичные фрагменты) и несовершенные (фрагменты, близкие в определенном смысле), прямые и симметричные, следующие подряд и разнесенные, индивидуальные (представленные лишь в одном из текстов обрататываемой совокупности) и совместные (характерные для группы текстов). Алгоритмы получения повторов указанного вида и представления их в удобной форме (частотное и лексикографическое упорядочение, префиксное дерево), а также некоторые классификационные процедуры реализованы в виде пакета прикладных программ "СИМВОЛ" [I], предназначенного для обработки текстов значительной длины (порядка 10^5 символов).

Хотя пакет в целом ориентирован на задачи дешифровочного типа, в ряде случаев он может быть использован и для анализа структурированных текстов, каковыми являются тексты на естественном языке. Укажем соответствующие возможности.

I. Получение статистик словоформ и словосочетаний по исходному тексту либо его лемматизированному аналогу. Понятие несовершенного повтора может быть использовано при этом для поиска словосочетаний с переменным компонентом.

2. Выявление ошибок. Анализ коротких низкочастотных ℓ-грамм позволяет обнаруживать отдельные классы орфографических ошибок, не прибегая к словарю. Выделение в "скользящем окне" словоформ, близких к смысле редакционного расстояния, часто сигнализирует о наличии семантических ошибок, связанных с появлением паронимов, либо о стилистических огрехах. Анализ характера ветвления дерева ℓ-грамм при больших значениях ℓ позволяет обнаруживать отдельные классы синтаксических ошибок (ошибки согласования). Наличие сверх-

длинных повторов в спектре ℓ-грамм часто бывает связано с ошибками при наборе текста (повтор строки).

3. Стилеметрия, установление авторства. Важная различительная информация содержится не только в предпочтительном употреблении автором некоторого набора слов, но и в распределении частей речи по позициям внутри предложения, в сочетаемости (глубины 2,3 и более) различных частей речи друг с другом и т.п. [2]. Получение указанных характеристик заложено в ППП "СИМВОЛ".

4. Дополнительное структурирование текста. Часто возникает необходимость в выделении наряду с традиционными структурными элементами (слово, предложение, абзац) некоторых дополнительных (зоны связи слов в предложении, сверхфразовые единства и т.п.). Выделение устойчивых (повторяющихся) комбинаций значащих единиц текста на различных иерархических уровнях может быть положено в основу такого процесса.

5. Сжатие текстов и словарей. Сложностной анализ текста [3] является удобным инструментом для решения первой задачи; представление в виде дерева позволяет решить вторую. В обоих случаях сжатие основано на выявлении полного спектра повторов.

6. Комбинаторный анализ слов. Последний ориентирован на создание нетрадиционных словарей (словарь паронимов, палиндромов и др.). Эти словари помогают лучше осознать сам процесс словообразования и могут найти порой неожиданные применения. Словарь паронимов, к примеру, может быть использован для подбора рифм, обнаружения семантических ошибок, поиска оборотов, характеризующихся определенной "игрой слов" и т.п.

Л И Т Е Р А Т У Р А

I. Гусев В.Д., Косарев Ю.Г., Тимофеева М.К. и др. Пакет прикладных программ для анализа произвольных символьных последовательностей значительной длины (СИМВОЛ) // Структурный анализ символьных последовательностей. - Новосибирск, 1984. - Вып. 101: - Вычислительные системы. - С. 3-21.

2. В. Фукс. По всем правилам искусства // Искусство и ЭВМ. - М.: Мир, 1975. - С. 405-412.

3. Lempel A., Ziv J. On the Complexity of Finite Sequences // IEEE Trans. on Inf. Th. - 1976. - Vol. IT-22, N I. - P. 75-81.

БЕЛЯЕВ Б. М., КАЗАКОВ М. И.,
СУРКИС А. С., ЯБЛОНСКИЙ С. А.

СИСТЕМА ВЕДЕНИЯ И ОБРАБОТКИ СЛОВАРЕЙ ПРОЦЕССОРА РУССКОГО ЯЗЫКА РУССИКОН-1

Автоматизация трудоемких процессов ведения и обработки машинных словарей русского языка большой емкости является одной из центральных проблем при создании языковых процессоров [1].

В настоящей работе рассматриваются основные характеристики системы ведения и обработки словарей, входящей в процессор русского языка РУССИКОН-1. Описываемая система включает:

- комплекс машинных словарей;
- подсистему поиска в словарях;
- подсистему составления словника новых слов;
- подсистему обработки и добавления в базовый словарь новых слов;
- подсистему конверсии и компрессии словарей.

1. Комплекс машинных словарей состоит из базового машинного словаря русского языка, словарей суффиксов (сочетаний суффиксов), псевдосуффиксов, префиксов, словообразательных классов (СлК), флективных классов (ФК) и др. Базовый словарь может иметь один из четырех "внешних" форматов словарной статьи: формат ВИНИТИ [1], формат "словоизменительная основа + номер ФК + длина словообразовательной основы + номер СлК", формат "словообразовательная основа + суффиксы + номер ФК + номер СлК", формат "префиксы + корень + номер СлК". Хранится базовый словарь во "внутреннем" компрессированном формате и реализован в двух вариантах: на основе словаря Засориной (40 тыс. словоизменительных основ) и на основе словаря Зализняка (100 тыс. словоизменительных основ).

2. Подсистема поиска в словарях позволяет реализовать следующие варианты поиска:

- по полной/неполной (в словоформу включены знаки *) словоформе;
- по полной/неполной словоизменительной или словообразовательной основе;
- по префиксу, корню, суффиксу, псевдосуффиксу;

- по номеру флективного класса;
- по номеру словообразовательного класса;
- поиск в словарях ФК, СлК, префиксов, суффиксов и т. п.

3. Подсистема составления словника новых слов позволяет составлять словники на основе морфологического анализа текстов, представленных в виде ASCII-файлов и выделения слов, отсутствующих в базовом словаре.

4. Подсистема обработки и добавления нового слова в базовый словарь позволяет автоматизировать следующие этапы:

- определение номера ФК, длин словообразовательной и словоизменительной основ обрабатываемой словоформы; при этом для контроля правильности выбора ФК и соответствующих длин основ распечатываются все варианты словоизменения слова для выбранного ФК и дополнительно проверяется наличие чередования букв в основе с помощью словарей и таблиц нормализации основ с чередованием букв; новые варианты подстановки букв в основе заносятся в словари подстановок, а для "слов-исключений" все словоформы помещаются в словарь неизменяемых слов.
- выделение префиксов производится путем наложения префиксов из соответствующего словаря и проверки правильности выделения корня, для которого строятся все варианты словоизменения; дополнительно предоставляется возможность создания всех вариантов соединения приставок с одним корнем с одновременной проверкой возвратности и совместимости с суффиксами путем распечатки всех вариантов новых словоформ;
- определение словообразовательного класса по суффиксам основы; для контроля правильности для каждого из возможных СлК производится распечатка всех вариантов словоформ;

5. Подсистема конверсии и компрессии словарей позволяет осуществлять конверсию словарных статей базового словаря из "внутреннего" формата в один из четырех "внешних" форматов и обратно. Для компрессии статей словаря во внутреннем формате используются специальные пятибитовые коды и коды Хаффмэна.

Система реализована на языке СИ в среде MS-DOS. В ней широко используется многооконный интерфейс, что существенно облегчает работу лингвиста - обработчика словаря.

ЛИТЕРАТУРА

1. Итоги науки и техники / сер. Информатика, т. 8 - М.: ВИНИТИ 1984

БЕЛЯЕВА Л.Н.

ЛЕКСИКОГРАФИЧЕСКИЕ ПРОБЛЕМЫ СОЗДАНИЯ АВТОМАТИЧЕСКИХ СЛОВАРЕЙ СИСТЕМ МАШИННОГО ПЕРЕВОДА

Автоматические словари (АС) являются ядерной частью лингвистического обеспечения различных систем, ориентированных на автоматическую переработку текстов (АПТ) на естественном языке. Информация, включаемая в словарные статьи (СлСт) АС, зависит от ориентации и уровня реализации (промышленный, экспериментальный) соответствующей системы АПТ.

Основные проблемы, возникающие при создании АС для практической системы машинного перевода (МП), определяются

I. Выбором заглавной единицы СлСт. Заглавная единица выбирается таким образом, чтобы минимизировать объем словаря при оптимизации процедуры лексического и морфологического анализа текстов и времени этого анализа. Соответственно, единицами АС являются машинные основы и словоформы, а в структуру словарного обеспечения включаются специальные блоки машинной морфологии. При этом, машинная морфология является базой для анализа текстовых слов и, одновременно, базой расширения словаря для ряда языков с агглютинативным типом словообразования. Дело в том, что задача минимизации времени анализа приводит к необходимости введения нескольких регулярных основ для одной лексической единицы (ЛЕ). В этих случаях пополнение словаря вариантными СлСт может осуществляться программно от единой базовой СлСт. Такие процедуры целесообразны, например, для систем морфологического анализа финского и арабского языков.

2. Бинарностью практических систем МП. АС ориентированы на конкретную пару языков. Даже в случае создания многоязычных "суперсистем" МП, каждая подсистема ориентирована на одну пару языков.

Эта особенность АС определяет необходимость введения в структуру СлСт лексико-грамматических характеристик входной ЛЕ, модифицируемых по результатам морфологического анализа, так и лексико-грамматических характеристик выходной ЛЕ, необходимых для синтеза предложения. Естественно, эта бинарность касается лексико-грамматических характеристик, а семантические словари, особенно словари семантических метаязыков, составляющие в некоторых системах глубинный уровень описания, могут быть независимыми от пары языков и составлять постоянную часть системы МП.

3. Прагматической ориентацией систем МП. АС любых систем МП также жестко ориентированы на конкретные предметные области и типы текстов.

Эта ориентация определяет особый подход к отбору ЛЕ в состав АС: при создании АС предварительно осуществляется анализ достаточно больших корпусов текстов по заданной тематике. В ряде систем для формирования АС используются уже имеющиеся в системе АПТ тексты или уже созданные словари информационных систем.

Вообще, следует отметить, что с расширением функций информационных систем и включением в них МП, архивы этих систем, сгруппированных по определенным тематикам, являются оптимальным источником для отбора лексики в АС.

Ориентация на конкретную предметную область (ПО) является важной характеристикой АС системы МП, так как позволяет снимать на уровне лексического анализа многозначность лексических единиц и стандартизировать перевод терминологии. Соответственно, особые проблемы создания АС связаны с

4. Специфичностью описания значения слова. Информация о переводе является особым уровнем описания ЛЕ в словарной статье АС. В отличие от традиционных словарей из набора переводов в АС системы МП исключаются синонимы. Эта же особенность отличает эти АС от словарей других систем АПТ: СлСт экспертных и информационных систем могут включать указания синонимических и гипонимических связей.

Выбор переводного эквивалента в АС представляет собой сложную лингвистическую задачу, особенности решения которой определяются

- необходимостью минимизации набора переводов;
- особенностями анализа неопознанных слов;
- стандартизацией перевода терминов.

Информация о переводе может вводиться в словарную статью двумя способами:

- непосредственно в виде набора переводных эквивалентов, задаваемых основой и морфологической характеристикой;
- в виде отсылок к СлСт общего выходного АС, если такой выделен в системы МП.

АС систем МП являются гибкими и динамически развиваемыми программными средствами их создания, ведения, пополнения и коррекции. Модульность автоматического словарного обеспечения позволяет поддерживать словарь в актуальном состоянии без его коренной реорганизации.

БЕРГЕЛЬСОН М.Б.

ФОРМИРОВАНИЕ ПРОПОЗИЦИОНАЛЬНОЙ СТРУКТУРЫ РЕПЛИКИ В ДИАЛОГЕ

Особенностью формирования пропозициональной структуры (ПС) в диалоге является то, что обычно стимулом для говорящего служит уже вербализованная информация, полученная из предыдущей реплики.

Иллокутивный тип предыдущей реплики влияет на способ (маршрут) вхождения в базу данных при синтезе ПС следующей реплики. Базой данных здесь служит словарь обобщенных ситуаций, представляющий собой тезаурус лексико-семантических групп русских глаголов.

Понятие обобщенной ситуации связано с довербальным представлением-события; наиболее близким аналогом выступает фреймовая структура ситуации.

Факторами, влияющими на процедуру формирования ПС, являются представления говорящего о ролевой структуре данной обобщенной ситуации; о коммуникативной значимости отдельных фрагментов ситуации и ее участников и распределении их по соответствующим позициям; о лексико-синтаксических штампах, связанных с данной обобщенной ситуацией.

Ролевая структура, зависящая, в первую очередь, от иерархии активности участников ситуации, определяет выбор соответствующей глагольной лексемы с присущей ей исходной синтаксической структурой. Последняя подвергается деривационным преобразованиям в зависимости от иерархии коммуникативной значимости.

Коммуникативная значимость элемента структуры определяется степенью его близости к фокусу внимания и/или к информационному фокусу.

Лексико-синтаксические штампы играют в процедуре формирования ПС роль своего рода "прерываний". Будучи активированы через некоторую лексему, они вызывают уже готовую синтаксическую структуру, которая воспринимается как некий синтетический объект.

Ясно, что для синтеза ПС описание самой процедуры менее важно, чем описание структуры базы данных, одним из главных свойств которой должна быть взаимозависимость информации в различных ее полях и возможность параллельного и одновременного обращения к различным типам информации.

БОРКОВСКАЯ Л.М., ЛЕВИЦКИЙ В.В.

СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ ОТНОШЕНИЙ МЕЖДУ ЛЕКСИКОЙ И СИНТАКСИСОМ

Целью работы является исследование взаимозависимости между типом придаточного предложения и его лексическим наполнением. Если в предложении выражаются локальные, темпоральные и тому подобные отношения,то не окажется ли так, что определенным синтаксическим типам,задаваемым, например, союзом,соответствуют определенные семантические "наполнители" ?

Материалом исследования послужила современная немецкая художественная проза общим объемом 350 тыс. словоупотреблений. Для статистической обработки фиксировались: а/тип придаточного предложения; б/глаголы,употребленные в роли сказуемого в обеих частях сложного предложения.Полученный корпус глаголов был разбит на 20 семантических подклассов.

Для выявления соотношений между типом предложений и семантическим подклассом глагола использовался критерий хи-квадрат и коэффициенты сопряженности.

Полученные результаты со всей очевидностью свидетельствуют о том, что определенным синтаксическим единицам, равным структурным частям сложно-подчиненного предложения /главному или придаточному/,соответствуют определенные классы лексических наполнителей. Так,например, обнаружена взаимосвязь глаголов движения с предложениями места и времени, глаголов передачи, получения и наличия - с предложениями условия и т.п. Эти предварительные данные дают основание предполагать, что между лексикой и синтаксисом существуют закономерные связи, скрытые от поверхностного наблюдения и могущие быть обнаруженными лишь с помощью специальных статистических приемов.

Дальнейшие разыскания в области межуровневых отношений следует направить на исследование других типов синтаксических конструкций и лексических наполнений. Очевидно, определенным типам глаголов в том или ином предложении должны соответствовать определенные типы наречий и существительных. Таким образом, требует выяснения система взаимосвязей всех наполнителей главных и придаточных предложений. Такого рода исследования, проведенные в сопоставительном плане /на материале различных языков/,помогут ответить на вопрос,какие типы взаимосвязей обусловлены внеязыковыми, а какие -внутриязыковыми факторами.

БРАТЫЩЕНКО Л.Г.

АВТОМАТИЧЕСКИЙ АНАЛИЗ СОЮЗНЫХ СОЧИНИТЕЛЬНЫХ РЯДОВ

0. Алгоритм анализа союзных сочинительных рядов является структурным блоком алгоритма автоматического выделения именных групп /ИГ/ в научном тексте.

I. При автоматическом распознавании именных групп особого анализа требует употребление сочинительных союзов. Цель анализа - ответ на вопрос: можно ли включать данный союз в ИГ, т.е. формирует ли он сочинительный ряд именных классов?

2. Лексический состав союзных сочинительных рядов в пределах ИГ ограничивается союзами "и", "или", которые не требуют употребления запятой /при автоматическом выделении именных групп в тексте запятая исключается из структуры ИГ/.

3. Союзы "и", "или" отличаются многообразием своих функций в тексте: участвуют в разделении частей сложного предложения, употребляются при перечислении однородных членов любых морфологических классов, используются для смыслового подчеркивания отдельных слов и т.д. Но в рамках ИГ реализуется только одна функция союзов - объединение однородных членов, причем однородными здесь оказываются только слова именных классов.

4. По отношению к левому и правому окружению союзов устанавливаются признаки однородности, которые предназначены для формального распознавания однородных членов. В соответствии с данными признаками объединение левого и правого окружения в один союзный сочинительный ряд происходит в следующих ситуациях: а/"...даны результаты тензометрических и рентгеновских измерений..."; б/"...исследуются возможности спроса и предложений..."; в/"...вводится описание геометрических тел и геометрических операндов..."; г/"...относящиеся к методам доступа НДАМ и НІДАМ..."; д/"...рассматривается простая система общения человека и ЭВМ на естественном языке..." и др.

5. Признаки однородности служат ограничениями на структуру ИГ с союзами. Благодаря этому удается избежать образования ошибочных сочинительных рядов, оставляя союз вне границ ИГ в таких случаях, как: "...вводятся коэффициенты для сравниваемых выражений и подсчитываются..."; "...предусматривают использование ЭВМ и для нечисленных алгебраических задач..."; "...обеспечивает прямой доступ к данным на магнитных лентах и обмен данными с магнитным барабаном..." и т.п.

БРЕЙДО Е.М.

МОДЕЛЬ СТИХОТВОРЕНИЯ /РЕАЛИЗАЦИЯ НА ПЭВМ/

Рассматриваемая ниже модель включает следующие параметры поэтического текста:

1. Число строк в стихотворении.
2. Размер.
3. Ритм.
4. Рифму.
5. Неотмеченные сочетания и различные грамматические конструкции с такими сочетаниями.

Алгоритм построен следующим образом. В машину введён словарь из нескольких сотен слов различного числа, рода, времени – в данном случае слова взяты из сб. Tristia О.Мандельштама. Словарь состоит из 4-х разделов: существительные и местоимения любого рода и числа /являющиеся в предложении подлежащими/, прилагательные и притяжательные местоимения любого рода и числа /служащие в предложении определениями/, глаголы различного времени, рода и числа /являющиеся в предложении сказуемыми/, существительные с предлогами, которые служат в предложении обстоятельствами. Каждое слово сопровождается информацией о метрике, рифме, грамматике и семантическом поле, к которому оно относится /если такая информация есть/. Для определения сочетаемости слов задана таблица сочетаемости семантических полей. Метрическая информация задаётся 2-мя числами – количеством слогов до ударения – $S_{до}$ и количеством слогов после ударения – $S_{после}$. Грамматическая информация задаётся указанием того, каким членом предложения может служить данное слово, и значениями 3-х параметров слова – рода, числа, времени. Кроме обычных значений, каждый из этих параметров может принимать ещё безразличное /нулевое/ значение. Например, безразличным является время у существительных и прилагательных. Информация о рифме задаётся в словах с ударением на последнем слоге /мужская рифма/, на предпоследнем /женская/ и на 3-м от конца /дактилическая/. Точная рифма определяется по словарю сочетаемости гласных и согласных, а неточная – по словарю неточных рифм.

Получаемое машиной задание включает:
1. Желаемое число строк в строфе и количество строф.
2. Тип рифмы /если она неточная/ и способ рифмовки. По умолчанию рифма подразумевается точной.
3. Размер /с уточняющими характеристиками/ и число слогов, отклоняющихся от заданного размера, а также номер слога, после которого идёт отклонение. Таких отклонений может быть несколько в строке или не быть совсем.
4. Число неотмеченных сочетаний в "стихотворении" /в частности, может быть равно нулю/.

"Стихи" пишутся справа налево. Сначала подбираются последние слова во всех строчках строфы в соответствии с заданным типом рифмы и способом рифмовки. Все остальные слова приписываются слева. Они берутся наугад, но могут быть отвергнуты из метрических, грамматических или семантических соображений.

В настоящее время на ПЭВМ /IBM PC/ реализована часть изложенного выше алгоритма, включающая "написание стихов" с заданным числом строк, размером и ритмическими вариациями /размер и ритм задаются для каждой строки отдельно/. Программа представляет собой диалоговую систему, написанную на языке пакета FOXBASE+. Предполагается, что предложенная модель должна развиваться за счёт увеличения числа параметров.

Предлагаемая программная система может служить инструментом для решения следующих задач:
1. Моделирование параметров, свойственных поэтике рассматриваемого автора /что позволит отточить стиховедческий аппарат/.
2. Сопоставление по заданному набору параметров различных индивидуально-художественных систем.
3. Построение новых, не используемых на сегодняшний день в русской поэзии, конструкций стиха.

ВАКУЛЕНКО Н.С.

ДЕРЕВО ТРИАД КАК СРЕДСТВО ОПИСАНИЯ СИСТЕМНОЙ ОРГАНИЗАЦИИ ЯЗЫКА

Взгляд на язык как на объект компьютеризации ведет к попыткам извлечь из языка все, что поддается формализации, и использовать в той или иной прикладной области. Однако постоянно ощущается какой-то верхний предел в постижении языковых явлений. Выдвигаются разного рода соображения, пытающиеся показать, где кроются причины ограниченности и где искать дальнейшие пути новых стратегий. Нами предлагается еще один подход к этой проблеме и проект метаязыка, позволяющего описывать явления явления, содержащие неописываемый элемент. Конструктивность введения этого элемента в описание лингвистических явлений можно в какой-то мере уподобить применению "черного ящика" в науке и технике.

Толчком к появлению такой идеи послужил опыт преподавания английского языка на основе метода формализации, включая применение ЭВМ (автором создано 15 учебных программ, внедренных в учебный процесс ЛИАП), когда встал вопрос о месте формализма в языковой системе. Кроме того, преподавание, как особый вид общения, поднимает вопрос о лингвистическом обосновании этого процесса, в том числе и организации "общения" человек – ЭВМ при компьютеризации обучения.

Как ни одно живое явление, по-видимому, не формализуемо полностью, так и в языковой системе наблюдаются несистемные элементы и невозможно избежать антиномий, перед которыми существующие логики оказываются ограниченными.

Итак, исходной целью являлось обоснование средствами прикладной лингвистики методики преподавания иностранного языка. Главные объекты исследования – человек и язык, причем, человек как воспринимающий язык в формализованном виде: преподавание всегда есть результат какой-то степени формализации. Если человек для нас исчерпывается понятиями "стимул – реакция", то и методика будет соответствующего типа. Мы исходим из факта, что обучаемый (любого возраста) – это интеллектуал, следовательно цель обучения – понимание и освоение внутренних законов языка, а предмет – язык как аспект человека (нет человека без языка).

Кроме того, 1) исчерпывающее описание человека немыслимо, а описание по частям – нелепо; 2) описание должно быть в

единстве, но может быть в разных аспектах; 3) аналогичные утверждения принимаются для языка. Далее, исходя из положения Соссюра о системности языка, мы утверждаем, что I) система языка описывает весь язык; 2) в каждом элементе языка заложен весь язык. Эти и другие соображения привели к мысли о необходимости упорядочить знания о явлениях языка, что могло бы внести вклад в такие сферы компьютерной лингвистики, как создание БД, искусственный интеллект, автоматический анализ текста, компьютерное обучение и др.

Предлагаемый метаязык состоит из элементов - иерархических триад, которые с точки зрения синтаксиса делятся на трады-элементы и триады-связи, а семантически - на триады иерархии, методологические, наложения, связи, развертки/свертки (внутреннюю, внешнюю). Основополагающими являются триады иерархии и методологические. Грамматикой данного языка порождается дерево триад, корнем которого является триада "человек". Путем наложения и свертки дерево может быть превращено в цепочку триад (в пределе - одну триаду). Порождение дерева осуществляется функциями внешней развертки и компарации. Глобальным критерием истинности является ОДНОСТОРОННЕЕ СИНТЕТИЧЕСКОЕ ТОЖДЕСТВО Я.Друскина.

Отличительными особенностями данного метаязыка являются: I) введение категории действительного, как попытка вынести в качестве неизвестного факт онтологичности явления, признавая онтологичное, во-первых, предметом веры каждого исследователя, во-вторых, систематизирующим фактором от макро- до микро-проявлений системности, полагая его принципиально неопределимым; 2) нарушение традиционных законов логики, как следствие I-го; 3) использование логики Я.Друскина приводит к формулированию парадокса, когда I=3=2, что доказывается рядом теорем; 4) возможность онтологической и гносеологической интерпретации явления системности языка; 5) включение элемента прагматики через аспектуализацию взгляда и ракурс аспектности; и др.

Разработана одна из ветвей дерева триад, приводящая к объяснению места и роли формализма в языковой системе как обоснование целесообразности использования его в системе преподавания иностранных языков.

ВИКС, Ю., ЛЕПП, М.

ИНСТРУМЕНТАЛЬНАЯ СИСТЕМА СОЗДАНИЯ БАЗЫ ДАННЫХ СЛОВАРЕЙ

Инструментальная система PROFUNDA.LEX предназначена для автоматизации процессов структурного анализа традиционных словарей и создания соответствующей базы данных, а также для синтеза новых словарей [1, 2]. Логическая структура системы PROFUNDA.LEX представлена на рисунке.

Подсистема ELMA - система генерирования процессоров словарей - базируется на инструментальной системе составления трансляторов для языков программирования [3].

Необходимой информацией для генерирования процессора словаря является формальное описание словаря. Формальное описание составляется пользователем при помощи диалоговой системы СЛОВАРЬ, спроектированной как совокупность типовых описаний словарей. В результате отредактирования образцов получается описание, которое соответствует словарю пользователя.

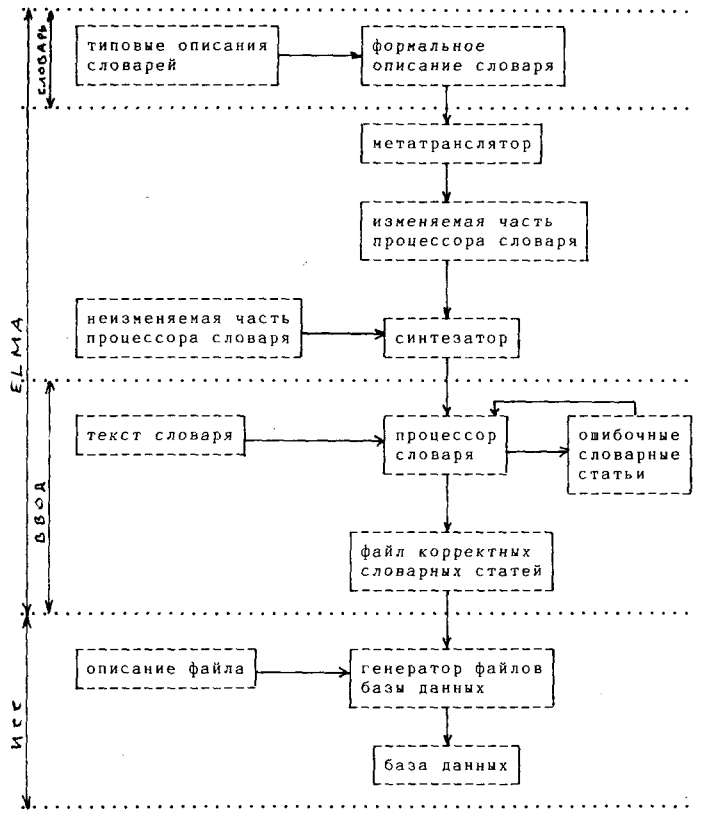
На основе получаемого метаязычного описания метатранслятор генерирует изменяемую часть процессора словаря.

Процессор словаря в целом синтезируется из двух компонентов - из изменяемой и неизменяемой частей процессора. Неизменяемая часть является инвариантным компонентом всех словарей и базируется на универсальных принципах структурного построения словарей. Изменяемая часть отражает специфику конкретного словаря и является прямым результатом формального описания словаря.

Процессор словаря работает как специфическая программа ввода данных, входом которого является последовательный текст словаря (с магнитного носителя или с терминала). Процессор словаря анализирует словарные статьи и формирует файлы корректных и ошибочных словарных статей. Из файла корректных статей образуется исходный файл базы данных, а файл ошибочных статей передается после редактирования снова процессору словаря.

Подсистема ИСС - информационная система словарей - представляет собой базу данных словарей (БДС) в виде совокупности файлов данных и описаний файлов [4].

БДС имеет модульную структуру, которая учитывает универсальные закономерности архитектуры словарей.



ЛИТЕРАТУРА

[1] Lepp M., Vooglaid A., Vykandu L., ELMA - an Instrumental Tool to build Programming Systems. - Informatique-85. Tallinn "Valgus" 1987, pp. 130-136.

[2] Õunapuu E., Infosüsteemide ganareerimise süsteem GENSI. Kasutamise juhend. TPI rotaprint. Tallinn 1988.

[3] Viks Ü., A Database of Dictionaries: why, what and how. - ALSAR 88. Tallinn 1990.

[4] Викс, Ю., Вход базы данных словарей в машинных фондах языков СССР. - Машинные фонды языков народов СССР. Материалы рабочего совещания (Таллинн, 19-22 декабря 1988). Таллинн 1988 (**A9**).

ГЕРМАН О.В., ГЕРМАНОВИЧ Е.И.

ПОСТРОЕНИЕ ГИПЕРТЕКСТОВЫХ СИСТЕМ НА БАЗЕ ЯЗЫКА СТРУКТУРНО-ФУНКЦИОНАЛЬНЫХ СПЕЦИФИКАЦИЙ

Описываются возможности языка новой ориентации, сменяющей парадигму программирования парадигмой проектирования и решения задач. Этот язык располагает набором нестандартных операторов, не имеющих аналогов в классических языках. К таким операторам относятся оператор выбора альтернативы (принятия решения), оператор создания альтернативы, оператор консультации, оператор возврата к контексту, оператор эвристики, оператор объяснения и др. Язык содержит два типа трансляторов: транслятор структуры программы, который создает граф программы, визуально отображаемый на экране ПЭВМ, и функциональный транслятор, транслирующий внешнее функциональное описание программы в множество выражений-кодов языка, связываемых с вершинами графа-программы. Язык "включает" пользователя в процесс обработки программы, предоставляя ему возможность логического выбора исполняемых вершин текущего графа программы.

При создании гипертекстовой системы структура программы кодирует логическую структуру текста. Последняя формально соответствует концепту

$$C = \langle C_1, C_2, \ldots, C_n \rangle,$$

где каждый C_j – либо новый концепт, либо лист. Лист соответствует фрагменту текста, для которого предусмотрены операции редактирования, поиска, синтеза и объединения с другими листьями. Вершина графа программы, не являющаяся листом, может быть вершиной-вопросом (Q), вершиной-альтернативой (A) или вершиной-условием (предикатом) (P).

Таким образом, логический маршрут из одной вершины в другую есть цепочка вида

$$\langle C_{i_0}, C_{i_1}, \ldots, C_{i_j} \rangle,$$

где $C_{i_k} \in \{Q, A, P\}$.

Операция поиска текста имеет формат

$$\langle C^*_{m_1}, C^*_{m_2}, \ldots, C^*_{m_g} \rangle \ find$$

где C_m – концепт, сопоставляемый вершине n_i, не являющейся листом.

Если концепт имеет тип Q, то пользователю предлагается ответить на вопрос; в случае вершины типа A – выбрать одну из

предлагаемых альтернатив; вершина типа P "рассчитывается автоматически" по имеющимся данным. Лист считается доступным (открытым), если указанные в операторе *find* концепты таковы, что:

- они связаны одним маршрутом с этим листом;
- все предикаты (условия) в *find* истинны (для этого маршрута.

Логическая организация текстов позволяет:

- строить на основе текстов системы консультаций, референтов, интеллектуальные справочники, системы поиска документов;
- для каждого текстового документа выделить его новизну, полезность, цель, применимость, область, путем создания логического фрейма (концепта) текста и его функциональной спецификации;
- формализовать перечень задач по работе с текстовыми документами, включив в процедуры обработки их семантику и прагматику.

Вмешательство человека в выполнение программы заменяет такие автоматические процессы в Прологе, как выбор альтернативы и бэктрекинг. Более того, пользователь вообще может установить такой контекст, который при выполнении программы не возникал. Основная задача языка - поддержка функций обработки графа решения в пространстве состояний задачи.

В докладе рассматриваются лингвистические, прагматические и философские аспекты выдвинутой парадигмы. Приводится спецификация и интерпретация "интеллектуальных" операторов, отсутствующих в традиционных языках программирования.

ГИНЗБУРГ Е.Л., ДВИНЯНИНОВА Г.С.

МЕЖКАТЕГОРИАЛЬНАЯ ОМОНИМИЯ РУССКИХ ПРЕДЛОГОВ
(опыт автоматизации лингвистического исследования)

Изучение материалов, отраженных в словарях современного русского литературного языка, позволило составить предлагаемое ниже табличное описание межкатегориальной омонимии предлогов.

Классы омоком-плексов	Части речи									Кол-во омо-комплексов в классе
	D	N	Cj	Pt	Prd	V	Intr.	Intj	Pstp.	
вблизи	+									67
край		+								2
включая						+				7
ради									+	I
в результате	+		+							2
вперед	+							+		I
круг	+	+								I
против	+				+					4
вроде			+	+						I
плюс		+	+							I
за	+	+			+					I
напротив	+			+			+			I

А. Отношение между частями речи сквозь призму омонимии с предлогами

I) Если в омонимический комплекс, кроме предлога входит междометие, то в него же входит наречие: Prp & Intj → D, например, вперед.

2) Если в омонимический комплекс, кроме предлога входит вводное слово, то в него же входит частица и наречие: Prp & Intr → Pt и D, например, напротив.

3) Если в омонимический комплекс, кроме предлога входит предикатив, то в него же входит наречие, или наречие и существительное: Prp & Prd → D или (D и N), например, против; за.

4) Если в омонимический комплекс, кроме предлога входит частица, то в него же входит наречие и союз: Prp & Pt → D или Cj, например, напротив; вроде.

5) Если в омонимический комплекс, кроме предлога входит союз, то в него же входит или наречие, или частица, или существительное: Prp & Cj → или D, или N, или Pt, например, в результате; вроде; плюс.

Б. Отношение между омонимическими комплексами

С учетом транзитивности импликаций, если существует омонимический комплекс, в который помимо предлога входит только

а) наречие (напр., вблизи), то существует омонимический комплекс, в который входит

(I) еще и союз, например, в результате;

(2) еще и междометие, например, вперед;

(3) еще и существительное, например, круг;

(4) еще и предикатив, например, против;

(5) еще и частица и одновременно вводное слово, например, напротив;

б) существительное (напр., край), то существует омонимический комплекс, в который входит

(I) еще и наречие, например, круг;

(2) еще и союз, например, плюс;

в) наречие и существительное (напр., круг), то существует омонимический комплекс, в который входит предикатив, например, за;

г) наречие и предикатив (напр., против), то существует омонимический комплекс, в который входит существительное, например, за.

С точки зрения этих правил фундаментальными для установления связей между классами омокомплексов должны быть признаны существительное, наречие, предикатив.

Выводы

Соизмеримость числа импликативных правил с числом классов, объединяемых этими правилами, свидетельствует о слабой логической организации омонимических комплексов. Если и говорить об организации этой части словаря, то лишь как о вероятностной, для которой определяющим должно признать наличие атракторов таких классов омонимических комплексов, мощность которых существенно превышает мощность других классов омокомплексов. По мощности классы омокомплексов упорядочены так, что их представители могут быть представлены последовательностью: тип вблизи, тип включая, тип против, тип край и тип в результате, тип ради и тип вперед, а также типы круг, вроде, плюс, за, напротив. В этой последовательности резко отличаются классы-атракторы. Это тип включая, тип в результате, тип против.

ГИНЗБУРГ Е.Л., КАЛАШНИКОВ Э.В., МАРОН М.Е.

АЛГОРИТМ ОЦЕНКИ КОРРЕЛЯЦИИ ТЕКСТОВ ПО ЧАСТОТНЫМ СЛОВАРЯМ И ЕГО РЕАЛИЗАЦИЯ.

Работа посвящена приложениям известных математических методов количественных оценок корреляции к лингвистическим объектам , а также разработке алгоритмических процедур этих оценок.Исходные тексты представляются в виде их частотных словарей.Каждое слово в своём словаре характеризуется набором параметров:грамматической категорией,абсолютной,относительной и накопленной частотами употребления в тексте,соответствующим рангом,а также коэффициентами рассеяния и употребительности.На первой ступени алгоритма анализируется общий лексический состав каждой пары словарей /текстов/ без учёта частоты употребления лексических единиц.Вторая ступень алгоритма учитывает совместные частоты употребления лексических единиц в пересекающихся подмножествах как по всему спектру текстовых полей,так и по их наиболее частотным фрагментам,а также грамматическим категориям и другим параметрам.На третьей ступени алгоритма,основываясь на результатах обработки данных на второй ступени,определяется степень квантитативной связи между словарями с помощью Спирменовской и Пирсоновской статистических моделей корреляции,основанных соответственно на частотном и ранговом аспектах.Алгоритм программно реализован на языке СИ,а его апробация проведена на ПЭВМ типа ИБМ.В качестве экспериментального лингвистического материала были использованы тексты двух циклов стихотворений Б.Пастернака общим объёмом около 6.000 словоупотреблений.С использованием программы лемматизации Ж.Мошкович по ним были составлены частотные словари и проведён сравнительный анализ в соответствии с рассмотренными выше алгоритмическими процедурами.В частности,были оценены:степень пересечения обоих словарей по всему лексическому полю текстов,по отдельным его фрагментам,а также словарному составу грамматических категорий.Экспериментальная проверка показала работоспособность предложенных процедур обработки д а н н ы х . Рассмотренный подход,по мнению авторов, может быть использован при решении проблемно ориентированных задач,возникающих в процессе создания,а также функционирования словарных и текстовых баз данных,насчитывающих большие объёмы лингвистической информации.

ГОЛОВНЯ А.И.

ПРОБЛЕМА ОМОНИМИИ НА ВХОДЕ И ВЫХОДЕ ЭВМ

Использование естественного языка в интеллектуальных системах ставит перед исследователями и разработчиками множество серьезных задач. Одной из таких задач является проблема омонимии, пронизывающей языковую систему на уровне словоформ и порождающей омонимию на уровне синтаксиса. В НИЛ теоретической и прикладной лингвистики БГУ им. В.И.Ленина теоретически установлено и доказано существование восьми типов омонимов: межклассовые, межпозиционные, межкорневые и пять их комбинаторных вариантов. Особенно мощно в русском языке представлена межпозиционная, или грамматическая, омонимия. Достаточно отметить, что почти регулярно не различаются позиции именительного и винительного падежей в мужском, женском и среднем роде большого числа существительных и прилагательных, что при обработке словоформы на входе создает определенные помехи и ставит задачу выбора кодировки словоформ.

Межкорневая, или лексическая, омонимия частично снимается при создании словарей предметной области тем, что определенная часть таких омонимов распределяется по разным словарям (ср. хлопок и хлопок, проказа "шалость" и проказа "болезнь", каток "место для катания" и каток "машина" и т.п.).

При снятии грамматической омонимии на входе единственным средством является анализ окружения омонимической словоформы и создание лакмус-контекстов, приводящих к однозначному выбору кодировки обрабатываемой словоформы. Эта задача прямо связана с созданием грамматического кода, нацеленного на снятие предполагаемых омонимичных ситуаций. Так, если в коде глагольного управления глаголов "видеть" и "стоять" будет упакована информация о том, что первый управляет винительным падежом, а в коде второго такой информации не зашито, то обращение к коду управления позволит в композиции "вижу стол" присвоить слову "стол" код винительного падежа, а в композиции "стоит стол" — код именительного падежа, что обеспечит на выходе правильное перекодирование и смысловую обработку.

Аналогичное кодирование прилагательных, предлогов, местоимений, числительных, учитывающее согласование, управление и примыкание как основные типы связей, — позволяет надеяться на успешное решение данной проблемы, находящейся на данном этапе в стадии научной проработки, что предполагает выявление всех типов омонимии на словаре объемом свыше 125 тысяч единиц, который будет надежной базой для снятия омонимии новых слов.

ГОЛЬДШТЕЙН С.Л., СЕВАСТЬЯНОВ А.А., ФУНШТЕЙН С.Л., ШЕРШНЕВ В.Н.

МОДЕЛИРОВАНИЕ ПРОФЕССИОНАЛЬНОГО ЯЗЫКОВОГО СОЗНАНИЯ МЕТОДОМ ПРАГМАТИЧЕСКОГО ВЗВЕШИВАНИЯ

Реализация систем ситуационно-прагматического управления ставит задачу описания пространства профессионального языкового сознания лиц принимающих решения (далее - ЛПР). С позиции метода семантического дифференциала (далее - СД) отмеченное пространство можно рассматривать как частную форму универсального трехфакторного пространства Ч.Осгуда.

Как отмечает А.Г.Шмелев, в частных формах СД общие факторы ОЦЕНКА, СИЛА, АКТИВНОСТЬ "наполняются определенным денотативным содержанием и видоизменяются структурно (объединяются, расщепляются и т.п.)" [I, с.9].

На наш взгляд, описание пространства профессионального языкового сознания ЛПР посредством методики СД затруднено тем, что набор шкал, предлагаемый исследователями может быть не актуален для управляемого объекта с позиции ЛПР. В этой связи предлагается метод прагматического взвешивания. Суть его в том, что эксперт должен последовательно разместить объекты относительно друг друга на абстрактной шкале. Когда в одной точке шкалы сталкиваются два различных объекта, "мерность" пространства увеличивается и появляется вторая ортогональная шкала. С ее помощью эксперт разводит два столкнувшихся объекта, а затем помещает на нее предыдущие объекты, используя расположение известных двух в качестве критерия. Аналогично могут появиться третья, четвертая и др. оси. По мнению авторов, в большинстве случаев размерность прагматических пространств должна ограничиться пятью-шестью осями. Данные денотативного (физического) описания объектов соотносятся с результатами прагматического взвешивания. На материале выявленных корреляций оси наполняются конкретным предметным смыслом. Не исключена ситуация, когда одна или несколько осей так и останутся свободными, что является указанием на недостаточность денотативных представлений об объекте. Предлагаемый метод может быть применим для создания микромоделей и нового типа экспертных систем.

I. Шмелев А.Г. Введение в экспериментальную психосемантику: теоретико-методологические основания и психодиагностические возможности.-М.:МГУ, 1983, 157 с.

ГОЛЬДШТЕЙН С.Л., СЕВАСТЬЯНОВ А.А., ФУНШТЕЙН С.Г., ШЕРШНЕВ В.Н.

СИТУАЦИОННО-ПРАГМАТИЧЕСКИЙ ПОДХОД К НЕКОТОРЫМ ЗАДАЧАМ ИИ

Одной из наиболее интересных областей применения "искусственного интеллекта" (ИИ) являются задачи управления сложными системами, включающими в качестве важного компонента человеческую деятельность.

Подобные системы можно разделить на две составляющие: субъект (лицо, принимающее решение) и объект (подсистема, на оптимизацию функционирования которой направлена деятельность субъекта). Задачей является передача функций субъекта системе управления. В существующих подходах системы управления строятся на основе моделирования функционирования объекта (математического или семантического) и моделирования процедуры принятия решения (оптимальное управление или логический вывод). Такой подход почти полностью ориентирован на объект и не учитывает свойства субъекта.

Несмотря на продуктивность данного подхода для ряда задач, отметим его принципиальные ограничения:

-чаще всего сложную систему не удается представить в виде замкнутой математической или семантической модели;

-как правило, цели управления лежат вне объекта и не поддаются формализации.

Очевидно также, что схема управления, по которой действует субъект мало похожа на схемы оптимального управления или логического вывода. По мнению авторов возможны системы управления, в которых бы главным образом моделировался не объект, а субъект.

Основные принципы построения таких систем:

-состояние объекта описывается набором его различных характеристик (пространство состояний объекта);

-субъект представлен профессиональным языковым сознанием в виде прагматического пространства, ориентированного на объект (на предметную область);

-между пространствами субъекта и объекта на основе экспертных данных устанавливается функциональная связь ("образ" объекта в пространстве субъекта);

-оценка ситуаций, принятие решений и их оптимизация проводятся в пространстве субъекта.

ГРЯЗНУХИНА Т.А., ДАРЧУК Н.П.

СИНТАКСИЧЕСКИЙ АНАЛИЗ В СИСТЕМЕ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ НАУЧНЫХ ТЕКСТОВ

Описываемая система автоматического синтаксического анализа /АСА/ создается в Институте языковедения АН УССР.

В задачи АСА входит получение информации о синтаксических связях двух уровней: а/ между словами в пределах простого предложения или предикативной части сложного; б/ между предикативными частями.

На уровне связей слов различаются координация, подчинение и сочинение.

Связи между предикативными частями представлены тремя основными типами: бессоюзной связью, сочинением, подчинением и шестью подтипами последнего. Дифференциация подчинительных связей зависит от средств, используемых для их организации. Выделяются: 1/ подчинение главной частью придаточной с помощью простого союза; 2/ союзного слова; 3/ парного расчлененного союза; 4/ союза и коррелятива в главной части; 5/ союзного слова и коррелятива; 6/ связь через соотносительные слова в обеих частях.

Связь между предикативными частями внутри сложного устанавливается через их предикативные центры /сказуемые/ и указанные союзные средства.

Исходными для алгоритма СА являются: а/ информация о словоформах текста, полученная в результате МА /код класса и код подкласса, репрезентирующий словоизменительные характеристики словоформы/, б/ порядок следования словоформ, в/ эмпирически выведенные текстовые комбинаторные свойства грамматических признаков определенных классов слов, г/ частичная семантическая информация некоторых групп слов.

Предшествующий синтаксическому анализу контекстный анализ для большинства слов текста снимает омонимию на уровне классов и для 85% – на уровне подклассов, устанавливает, по сути, почти все атрибутивные связи. Это в значительной мере облегчает проведение СА.

СА начинается с предварительного членения фразы на сегменты и одновременного выделения в них вводных слов и конструкций, которые исключаются из синтаксического представления предложения. Сегмент при этом определяется как последовательность словоформ с информацией к ним в виде двухэлементных кодов, ограни-

ченная знаками препинания /запятой, двоеточием, точкой с запятой, точкой/ и союзами. При этом элемент, разграничивающий два сегмента, относится к правому сегменту. Сегмент может оказаться осмысленным отрезком текста /простым предложением или предикативной частью сложного/, но может быть бессмысленным или незаконченным отрывком, состоящим, например, из одного слова.

Синтаксический анализ выполняется в два этапа: I - установление предикативных частей и определение характера связи между ними; II - анализ связей слов в пределах предикативных частей. Первый этап, в свою очередь, распадается на внутрисегментный и межсегментный.

В задачи внутрисегментного анализа входит определение в каждом сегменте явных компонентов предикативного центра или претендентов на их роль на основе грамматической информации, содержащейся в кодах классов и подклассов словоформ.

Целью межсегментного анализа является объединение сегментов в предикативные части. Исходной информацией при этом служат результаты внутрисегментного анализа: учитывается взаиморасположение сегментов с различными структурными типами предикативных центров /предикативная пара со сказуемым в препозиции, предикативная пара с подлежащим или претендентом на подлежащее, центр из одного только подлежащего или одного его претендента, центр из одного сказуемого или претендента на сказуемое/. На этом этапе анализа привлекается также информация о наличии в сегменте определенных союзов, местоимений и наречий.

Одновременно с установлением межсегментных связей происходит и снятие грамматической неоднозначности кодов тех слов, с которыми работают правила этих этапов.

В отдельных случаях допускается несколько вариантов определения границ предикативных частей в составе сложного.

В ходе определения синтаксических связей слов внутри предикативных частей происходит выделение обособленных оборотов, а также сравнительных конструкций.

Результаты анализа предлагается представлять в виде таблиц, в которых задаются пары номеров слов предложения, связанных между собой определенным синтаксическим отношением. Позиции номеров слов указывают направление связи. При этом исключается возможность занимать разные позиции одним и тем же словом.

ДАНИЛЬЧЕНКО К.С., МОСКАЛЕНКО Т.А., ПОЛИКАРПОВ А.А.

АВТОМАТИЗИРОВАННОЕ РАБОЧЕЕ МЕСТО ЛЕКСИКОЛОГА И ЛЕКСИКОГРАФА

1.АРМ-ЛЛ проектируется и создается с целью повышения эффективности лексикологических исследований.АРМ-ЛЛ позволяет фиксировать в интерактивном режиме взаимодействия с компьютером результаты интуитивной работы лингвиста с текстовым материалом (например,семантизация текста) и осуществляет алгоритмизированные процедуры обработки изучаемых текстов или лексикологических материалов (например,создание различных типов конкордансов текстов,словарей,словников,подсловарей).

2.Специальной целью создания АРМ-Л является автоматизация процесса создания автоматических семантических-частотных словарей (ЧС) следующих видов:

1.ЧС лексико-семантических вариантов(ЛСВ) с фиксацией частоты лексемы в целом и фиксацией частот ее словоформ,в т.ч. и частот их реализации в тех или иных ЛСВ.

2.ЧС синонимических групп ЛСВ.

3.ЧС синонимичных лексем.

4.ЧС гиперлексем.

5.ЧС ЛСВ гиперлексем.

6.ЧС синонимических групп гиперлексем.

7.ЧС синонимичных гиперлексем.

8.Комплексный ЧС семантических категорий ЛСВ,лексем, гиперлексем (с указанием вложимости категорий различных уровней в категории надуровней).

Все вышеприведенные виды словарей создаются на базе

совокупности текстов, отобранных лингвистом-пользователем. Тексты обрабатываются АРМ-ЛЛ с использованием информации из ряда вспомогательных словарей,поддерживающих контрольную и справочную функцию АРМ-ЛЛ.Речь идет о контроле вводимых текстов в отношении орфографической правильности, автоматическом выделении в тексте омонимов и омонимичных словоформ различных лексем.Кроме того,АРМ-ЛЛ осуществляет консультативно-диагностическую функцию используя дефинитивный словарь лексем русского языка на машинном носителе. Следовательно,используются следующие автоматические словари:а)словарь лексем современного русского языка;б)словарь омонимов; в)словарь омонимичных словоформ различных лексем; г)словрь гиперлексем.

Все словари, упомянутые выше в качестве вспомога тельных, предполагают наличие соответствующих программ ввода и коррекции словарных статей.При этом,с одной стороны,изучаемые тексты снабжаются информацией в форме помет,подсказок со стороны АРМ-ЛЛ,с другой стороны,словари пополняются и корректируются по мере ввода в систему новых текстов и обнаружения новых лексем и словоформ,а также неполноты существующих словарных статей.Процесс коррекции словарей предполагает многократный контроль предлагаемого оператором-лингвистом изменения (пополнения) словарной статьи и завершается принятием решения администратором АРМ-ЛЛ об окончательном включении предлагаемого изменения в словарь.

ДАРЧУК Н.П.

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ОШИБОК В ТЕКСТЕ

I. Каждый исследователь, занимающийся вопросами речевой деятельности, сталкивается с ее отклонениями - речевыми ошибками.

С внедрением ЭВМ новым источником ошибок /искажений/ является запись информации оператором, а не читающим автоматом. При наличии ошибок доступ к информации, к ее обработке, использованию сильно затруднен, а иногда невозможен. В связи с этим возникает проблема автокорректора, близкая к проблеме автоматизации корректуры в издательской работе.

Систематическое изучение номенклатуры ошибок может облегчить поиск причин возникновения ошибок, механизмов их образования, а самое главное - их коррекцию - процесс более сложный, чем обнаружение ошибок.

2. Задача автоматического обнаружения ошибок в тексте становится реальной при наличии иерархически организованного программного обеспечения автоматического анализа текста, нижнюю ступеньку которого составляют алгоритмы и программы морфологического анализа. Вследствие такого анализа возможно обнаружение орфографических и некоторых грамматических ошибок. Результаты работы МА могут служить основой для программ синтаксического анализа, тогда реально обнаружение синтаксических и пунктуационных ошибок.

3. С помощью разработанного в отделе структурно-математической лингвистики АН УССР автоматического морфологического анализа создана система автоматического обнаружения орфографических ошибок, в основу которой положен автоматический орфографический /канонический/ словарь слов со сведенной парадигмой, автоматически составленный на материале безошибочных текстов, тематически близких к исследуемым.

4. Процедуры контекстного анализа будут использованы для автоматического обнаружения ошибок синтаксического характера, связанных с нарушением норм русского языка, касающихся организации подчинительных связей слов с тексте. Для этого в КА вводятся специальные фильтры в блоках, анализирующих согласованность грамматических подклассов существительных и их определений, предлогов и управляемых ими именных форм, а также глаголов и подчиненных им форм существительных.

5. В основу системы обнаружения синтаксических и пунктуационных ошибок будет положен автоматический синтаксический анализ. Планируется завершить создание автокорректора в 1991 году.

ДАРЧУК Н.П., СОЛОМАНЧУК К.С.

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ГЛАГОЛЬНЫХ БЕСПРЕДЛОЖНЫХ СВЯЗЕЙ

Автоматическое установление глагольных связей представляется важной и достаточно сложной задачей, так как эта группа словосочетаний является самой многочисленной благодаря богатым валентностным свойствам глагола.

Поскольку каждый тип связи выражается определенной моделью сочетаемости, имеющей как структурные, так и позиционные варианты, при составлении алгоритма использовались сведения о текстовых характеристиках глагольных связей /расположение в пре- или постпозиции зависимого элемента относительно глагола/, данные о частоте модели и ее вариантах, почерпнутые из / I /.

Алгоритм построен по принципу одновременного выделения всех встретившихся в предложении моделей и состоит из таких основных частей: I. Поиск глагольного класса. 2. Поиск слов, грамматические характеристики которых соответствуют характеристикам компонентов заданных моделей. 3. Формирование подграфа глагольных связей. Поиск глагольного класса осуществляется по кодам, которые заранее присваиваются по АМА. В некоторых глагольных моделях учитывается информация, содержащаяся в подклассах глаголов /признак залога, инфинитива/.

Поиск связей глагола осуществляется в пределах простого предложения или каждой предикативной части сложного. По предварительным данным о частоте модели и частоте право- и левосторонних связей анализ рационально начинать с поиска правосторонних связей /их 86,4%/. Частота употребления моделей глагольных связей в тексте обусловливает следующую очередность проверок: I/ глагол+ наречие, 2/ глагол+имя$_{вин.п.}$, 3/ глагол+имя$_{дат.}$, 4/ глагол+имя$_{тв.}$, 5/ глагол+имя$_{род.}$, 6/ глагол+ аббревиатура/символ,формула/

В отдельную группу выделены правила анализа кратких причастий, кратких прилагательных и наречий, выполняющих в предложении функцию сказуемого. Их модели связей сужены до двух падежных форм ...+имя$_{дат.}$ и ...+имя$_{тв.}$

Если между зависимыми от глагола формами присутствует сочинительная связь, формальными выразителями которой являются союз и запятая, то последним приписываются номера тех элементов, которые они связывают.

ЛИТЕРАТУРА

I. Автоматизация анализа научного текста. – Киев:Наук.думка.1984. –257 С.

ДЕТЛОВС В.К.

МОТИВЫ И ЛАД

Свойства ступеней лада обнаруживаются как особенности геометрической формы мелодической линии вблизи данной ступени [3]. Рисунок мелодии является одним из средств музыкального формообразования. Поэтому можно искать непосредственные связи между микроединицами музыкальной формы (мотивами) и свойствами лада.

Составлен частотный словарь F-мотивов Бороды [1] для семисот латышских народных мелодий [2], учитывая ступень лада для опорного звука мотива. Опорным для простого такта T считается первый звук, а для возрастающей последовательности A последний звук. У полного мотива TA имеется два опорных звука. Подобный подход уже использовался в работах Бороды.

Анализ показал, что ямбические мотивы A составляют лишь 5% (см. также [4]); их опорные звуки отсутствуют на $\underline{\text{II}}$, $\underline{\text{V}}$, $\underline{\text{VI}}$ и $\underline{\text{VII}}$ ступенях при полном отсутствии их первых звуков на устойчивых ступенях $\underline{\text{I}}$, $\underline{\text{III}}$, $\underline{\text{V}}$. Мало и полных мотивов (9%). Первая опора реже всего встречается на ступенях $\underline{\text{II}}$ и $\underline{\text{V}}$, а вторая отсутствует на $\underline{\text{VI}}$ и $\underline{\text{VII}}$. Преобладают хореические мотивы T (59%), которых меньше на $\underline{\text{VI}}$ ступени. Оканчиваются же они реже всего на устойчивых ступенях.

С каждым мотивом можно связать его направление, понимаемое как интервал между первым и последним звуком. Оказалось, что в целом направления следуют закону о центральной и периферийной областях лада [3]. Очень любопытно, что среднее значение направления мотива по всем ступеням равно -0.239, т.е., сильно отличается от нуля.

ЛИТЕРАТУРА

1. Борода М.Г., Частотные структуры музыкальных текстов // Сб. статей. Тбилиси, 1977. с. 178-202.

2. Витолиньш Е.Я., Латышская народная музыка: Сватовские песни, Рига: Зинатне, 1986. 600с.

3. Детловс В.К., Лад и мелодическая линия // Алгебра и дискретная математика. Рига: ЛГУ, 1986. с. 67-87.

4. Детловс В.К., Моделирование хореических и ямбических интонаций латышских сватовских песен // Алгебра и дискретная математика. Рига: ЛГУ, 1989, с. 65-76.

ДЕТСКИНА Р.В., КАРПИЛОВИЧ Т.П.

НЕМЕЦКО-РУССКИЙ ПЕРЕВОД ЗАГОЛОВКОВ ПАТЕНТОВ НА БАЗЕ ПЭВМ

Один из способов получения сигнальной информации о текстах на различных языках - автоматизация анализа заголовочных конструкций. Заголовок - это основной признак, по которому пользователь определяет соответствие или несоответствие публикации своим профессиональным интересам. Разработка алгоритма немецко-русского машинного перевода (МП) на базе заголовков патентов потребовала решения таких задач, как выявление и формализация семантических структур заголовков, закономерностей их лексического и структурно-грамматического выражения, установление связей между заголовком и текстом.

Соотнесение заголовков и текстов рефератов показало, что в текстах обязательно присутствует заголовочная лексика. При этом отмечались такие способы повторных номинаций, как местоименные замены, словарная и контекстуальная синонимия, перифразы. Учитывались статистические данные о повторяемости компонентов заголовка в тексте реферата патента. Все это позволило сделать вывод, что в большинстве случаев заголовочная лексика несет в реферате значительную семантическую нагрузку, указывая на основные характеристики патентуемого метода или устройства.

Лингвистической информационной базой алгоритма МП является немецко-русский автоматический словарь словоформ по складскому хозяйству, типология базовых синтаксических моделей немецких заголовков и их русских переводов, набор диагностирующих признаков для распознавания выходной структуры заголовка. Доалгоритмический этап анализа показал, что исследуемые заголовочные конструкции носят именной характер. Основной единицей номинации в них является существительное, образующее в результате преили постпозитивного распространения именного ядра так называемые базовые синтаксические модели. Программная реализация алгоритма осуществлена на языке программирования PASCAL на базе персональной ЭВМ IBM PC/AT, операционной системы MS DOS . Для формирования словаря использована база данных dBASE III Plus, позволяющая оперативно изменять и пополнять словарь.

ДЖУБАНОВ А.Х.

ИСПОЛЬЗОВАНИЕ ЭВМ ДЛЯ ВЫДЕЛЕНИЯ ОПОРНЫХ СЛОВ ТЕКСТА

Проблемы выделения основного и факультативного содержания текста является одной из важных проблем инженерной лингвистики. Существует много различных способов выделения опорных (ключевых) слов. Наиболее перспективным нам представляется метод использованный в работах Марусенко М.А. и Зубова А.В. [1; 2]. В его основу положены идеи о том, что наиболее значимые единицы текста имеют наибольшую частоту употребления в тексте и встречаются в наибольшем количестве абзацев такого текста. Формально эта зависимость выражается в виде следующей формулы:

$$K_{вах} = \frac{F \cdot m}{N \cdot n}, \qquad (I)$$

где F – абсолютная частота слова с тексте, m – число абзацев, в которых встретилось слово, N – общее число слов в тексте, n – общее число абзацев в тексте.

Как видно, подсчеты по этой формуле требуют построения частотного словаря как отдельного текста, так и всех его абзацев. Ясно, что для достаточно больших текстов эта задача трудно выполнима.

Поэтому нами создана программа, которая строит частотные словари абзацев и в совокупности частотный словарь всего текста.

Предварительно перед вводом текста в ЭВМ все словоформы текста приводятся к канонической (словарной) форме. Затем машина строит частотные словари лексем каждого абзаца. И в заключении, компьютер, объединяя подобные словари абзацев строит частотно-алфавитный словарь всего текста. При этом для каждой лексемы словаря текста автоматически находится абсолютная частота лексемы в тексте (F, см. формулу (I)) и число абзацев, в которых встретилась данная лексема (m).

Все эти данные используются затем для автоматического отнесения слов анализируемого текста к числу главных или второстепенных опорных слов.

В докладе подробно на материале казахских текстов рассматривается процедура построения словарей абзацев и всего текста.

ЛИТЕРАТУРА

I. Зубов А.В. Статистический аспект содержания текста и его формальное представление // Квантитативная лингвистика и автоматический анализ текстов. – Тарту, 1986, с. 75-94.

2. Марусенко М.А. О формировании словника словаря статистически устойчивых научно-технических терминов // Структурная и прикладная лингвистика. – Л.: ЛГУ, 1983, с. 82-89.

ДМИТРИЕВА Н.В., ПОГОСЯНЦ А.Г.

ПРОГРАММА СЕМАНТИЧЕСКОГО СИНТЕЗА ЕДИНИЦ С ОБЩИМ ЗНАЧЕНИЕМ "ПЕРЕМЕЩЕНИЕ В ПРОСТРАНСТВЕ"

Особенность современного состояния лингвистической науки определяется необходимостью перехода от постановки и решения аналитических задач, конечной целью которых является описание и классификация исходного языкового материала, к формулированию и разрешению задач синтеза.

Построение действующей модели семантического синтеза и представление её в виде алгоритма образует реальную базу для создания программного продукта, решающего задачи лингвистической комбинаторики на уровне плана содержания и плана выражения.

При этом первоначально синтезируется содержательная структура, рассматриваемая как удовлетворяющая требованиям семантической достаточности и непротиворечивости комбинация семантических компонентов; далее производится операция синтеза на уровне плана выражения, - комбинированию подвергаются двусторонние единицы: морфемы и слова.

В результате пользователь получает информацию о всех возможных средствах выражения заданного им значения вне зависимости от уровневой принадлежности этих средств.

Программа семантического синтеза, обеспечивающая переход от сформированного в сознании продуцента речи смысла к структурно-организованному значению как факту лексико-семантической системы языка и от него - ко всем возможным средствам выражения данного значения, является функциональным аналогом процесса реального речепроизводства. Тем самым она не претендует на раскрытие истинных внутренних механизмов этого процесса, как он происходит в сознании говорящего, а лишь моделирует их.

В практическом плане сформулированная выше задача была поставлена и решена на материале семантического класса русских глаголов перемещения /около 150 непроизводных глаголов и более 1200 их дериватов/.

Центральным понятием, позволившим реализовать идею создания программы семантического синтеза глаголов с общим значением "перемещение в пространстве", стало понятие семантической модификации, под которой в наиболее общем виде понимается вслед за А.М. Пешковским "изменение значения на заданную величину".

На базе этого понятия в результате семасиологического анализа русских глаголов, образующих семантический класс глаголов пере-

мещения, был определён модификационный потенциал общего значения "перемещение в пространстве". К модификационным параметрам этого значения следует отнести параметры "Направленность", "Повторяемость", "Цель", "Скорость", "Время", "Расстояние", "Количество участников", "Каузация" /характеристика по факту распространения перемещения субъекта на другие тела/, "Сегментация" /возможность представления перемещения как цепи повторяющихся однотипных двигательных операций - сегментов/.

На этапе анализа устанавливаются системы дифференциальных признаков, актуальных для каждого из названных выше параметров, и определяются конкретные наборы выделяемых на основе этих ДП семантических компонентов.

Иерархия дифференциальных признаков во внутренней семантической структуре параметра является основой для создания серии вопросов, обращённых к пользователю, для идентификации значения, которое он стремится выразить.

Синтез содержательной структуры в программном продукте обеспечивается, таким образом, через интерфейс программы - диалог с пользователем. Для определения вхождения компонента каждого из параметров и его идентификации в составе заданной семантической структуры пользователю предлагается ряд сменяющих друг друга "деревьев" вопросов /каждое "дерево" вопросов соответствует иерархии ДП одного из параметров/. Вопросы предполагают либо ответ в форме "да - нет", либо выбор одного из предлагаемых вариантов ответа /режим меню/.

Для решения задачи синтеза средств выражения полученной комбинации используются полученные в ходе анализа результаты исследования возможностей выражения каждого из компонентов-модификаторов: они могут быть выражены: а/ супплетивно - совместно с модифицируемым значением - глагольным корнем /основой/, б/ самостоятельно, формально связанно с исходным глаголом - аффиксом, в/ самостоятельно, формально независимо от исходного глагола - наречием, предлогом.

На этапе синтеза средств выражения заданной семантической структуры программой осуществляется операция лингвистической комбинаторики имеющихся в языковой системе средств выражения отдельных компонентов. Алгоритм включает в себя информацию: а/ о морфонологических изменениях в процессе деривации, б/ об изменениях глагольной основы в процессе деривации и возможностях каждого из аффиксов присоединяться к основам различных типов, в/ о позициях самостоятельных лексических средств выражения компонентов-модификаторов относительно глагола при синтезе словосочетания.

ДОЛИНСКИЙ В.А.

ПОЛИСЕМИЯ В СЛОВАРЯХ И В АССОЦИАЦИЯХ

Данное исследование развивает представление об ассоциативном семантическом поле слова, как основе для представления его смысла в виде дискретного набора значений, или лексико-семантических вариантов, в толковых словарях. Это представление базируется на фундаментальном свойстве слов вызывать ассоциации с другими словами и исходит из признания ассоциативной природы разных типов или аспектов значения слова. Ассоциативные связи, существующие в сознании говорящих, потенциально бесконечны. Понимание смысла слова всегда связано с его оцениванием, с приданием разным участкам семантического поля, разным ассоциациям разных весов. С целью сопоставления весовых функций ассоциаций необходимо вводить нормировку - требовать, чтобы площадь, ограниченная весовой функцией и осью ассоциаций, была равна единице. С этим ограничением мы можем говорить уже о распределении вероятностей ассоциаций на семантическом поле, о вероятностном ассоциативном пространстве слова, или о мере потенциальной возможности раскрытия смысла слова. Ассоциативное значение - это значение, выявляемое посредством анализа дистрибуции ассоциаций, полученных от той или иной группы испытуемых в ответ на заданное слово-стимул в эксперименте, который можно планировать и повторять.

Свободный ассоциативный эксперимент является ценным источником информации при исследовании психических эквивалентов семантических полей. При выявлении связи ассоциативного значения слова в эксперименте и его полисемичности в толковом словаре, было обнаружено [I], что более многозначные слова вызывают ассоциации, более однородные /гомогенные/ по своему распределению. Особый интерес представляет возможность отслеживания динамики изменения ассоциативного значения слов с изменением возрастной характеристики групп испытуемых. Установлено, что с возрастом, в целом, понижается частота наиболее популярной ассоциации, возрастает ассортимент ассоциаций, снижается тангенс угла наклона распределений γ. Задачей данного анализа было выявить динамику расширения ассоциативных связей с возрастом испытуемых и степень количественного и качественного соответствия ассоциативных значений слов значениям, отраженным и представленным в лек-

сико-семантических вариантах словарей разного типа.

Материалом исследования послужили данные экспериментов по свободному ассоциированию с регистрацией первого ответа /дискретные ассоциации/, проводившихся с детьми 6-7 лет /группа Д/ и взрослыми /группа В/. Русский язык является родным для испытуемых. Эксперимент проводился индивидуально. Стимулы и реакции давались устно. Общее количество испытуемых - 400 /по 200 в каждой группе/. Число стимульных слов - I2. Полученные ассоциации были сгруппированы по ЛСВ, представленным в [3]. В таблице приводится индекс числа значений, выявленных в эксперименте, а также отраженных в трех толковых словарях: большом, среднем и малом.

Слова-стимулы	С Л О В А Р И [2]	[3]	[4]	Группа В	Группа Д
I Высокий	6	7	6	3	I
2 Низкий	8	6	7	4	2
3 Широкий	8	7	7	6	3
4 Узкий	6	4	4	3	I
5 Длинный	2	2	3	3	2
6 Короткий	4	3	4	3	2
7 Глубокий	4	5	5	4	2
8 Мелкий	6	7	6	7	3
9 Далекий	4	4	5	3	2
I0 Близкий	5	4	7	4	2
II Толстый	4	3	3	3	2
I2 Тонкий	II	I2	6	I2	4

ЛИТЕРАТУРА

I. Долинский В.А. Распределение реакций в экспериментах по вербальным ассоциациям. // Уч. зап. Тартуск. ун-та, вып. 827. - Тарту,I988. - С. 89-I0I.

2. Словарь современного русского литературного языка. В I7 тт. - М.,I948-I965.

3. Словарь русского языка. В 4 тт. - М.,I98I-I984.

4. Словарь русского языка /Ожегов С.И./. - М.,I983.

ДРИЗУЛЕ В.А.

ИСПОЛЬЗОВАНИЕ ЭВМ В СОЦИОЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Хотя необходимость использования электронно-вычислительной техники в области социальной лингвистики и вполне очевидна, тем не менее она чаще используется в демографических исследованиях. Несомненно, что это убыстрит проведение эксперимента и даст возможность получить более полные и достоверные результаты.

В Институте языка и литературы им.Андрея Упита Латвийской Академии Наук был собран достаточно представительный корпус анкет - более 1600 единиц - который затем был введен в память машины 1045 ЕС ЭВМ. Разработка анкет, сбор, их первичная индексация и ввод через дисплей осуществлялся силами лингвистов, разработка программы была осуществлена А.Генкиным, сотрудником ИЭ Латвийской АН.

В результате работы программы были получены основные статистические данные по всему массиву. Объектом анализа служили данные о фактическом функционировании как русского, так и латышского языков в республике. Значительно облегчается возможность получения дополнительных данных по пользовательским запросам, например, в масштабах только одного города или с учетом различных показателей.

Также введена в память ЭВМ оценка по определенной разработанной шкале лингвистических тестов, выполненных при заполнении анкет. Цель тестов - выявить интерферентные явления и оценить уровень знаний второго языка. Это также облегчит анализ и повысит информационную ценность исследования.

ЖИЛИНСКЕНЕ В., МАКСИМАВИЧЮТЕ Н., ПОЗНАНСКЕНЕ Д.

ПРОБЛЕМЫ КОМПЬЮТЕРИЗАЦИИ СЛОВАРЯ ЛИТОВСКОГО ЯЗЫКА

Объектом работы является толковый "Словарь современного литовского языка". Словарная статья однозначного слова состоит из: заглавного слова, его форм, ряда признаков, дефиниции, иллюстраций и дериватов от заглавного слова или от вышестоящего деривата. В статье многозначного слова приведены отдельные, состоящие из дефиниции и иллюстраций, значения, которые иногда имеют и свои признаки. В определенных статьях приведены фразеологизмы и соединения слов терминологического значения. Они в основном относятся к заглавному слову, иногда к отдельным значениям многозначного слова.

Технической базой автоматизации этого словаря является центральная ЭВМ - СМ-1420 и терминальная микро ЭВМ - ДВК-3. Используемое программное обеспечение: операционная система RSX-11M, система управления базой данных - АДАБАС-М.

Так как отдельные части словарной статьи очень длинные (до 1176 б), а длина поля АДАБАС-М не превышает 256 б, то проблема размещения данных в ЭВМ решается путем формирования 8 баз данных. I база данных охватывает словообразовательные части слова, II - признаки, дефиницию и иллюстрации слова или значения, III - фразеологизмы, IV - терминологические соединения слов, V, VI, VII, VIII базы данных назначены дериватам и их структура является аналогичной I-IV базам. Связывающей цепью между всеми этими базами служит заглавное слово.

В настоящее время реализовано:

1) поиск терминов по разным терминологическим признакам,

2) поиск заглавных слов или полных статей по диалектологическим и стилистическим признакам,

3) поиск всех слов заданной определенной части речи,

4) поиск слов как представителей определенной части речи, но с дополнительными признаками,

5) поиск слов по заданным определенным частям слова,

6) составление списков по акцентуационным парадигмам соответствующих частей речи.

Вопросы, связанные с детализацией вышеуказанных признаков и с программным обеспечением ввода данных словаря в ЭВМ, докладывались на конференции "Машинные фонды языков народов СССР", которая состоялась в Таллине в декабре 1988 г.

ЗАХАРОВ В.П., МУНТЯН Д.В., ПЕТРОВ А.Д.

ПРИНЦИПЫ ПОСТРОЕНИЯ ДИНАМИЧЕСКОГО АВТОМАТИЗИРОВАННОГО СЛОВАРЯ СЕМАНТИЧЕСКИ СВЯЗАННЫХ ТЕРМИНОВ

Построение почти любого словаря начинается с выбора и описания метода классификации лексики, предназначенной для включения в данный словарь. Таких методов может быть по крайней мере три:

1/ логико-интуитивный;

2/ метод ассоциативного эксперимента;

3/ дистрибутивно-статистический.

В основе всех вариантов последнего метода лежит анализ величин, характеризующих совместную встречаемость элементов текста в контекстах определенной длины. Предполагается, что слова, часто встречающиеся вместе в пределах того или иного интервала текста, связаны между собой семантически.

Поскольку дистрибутивно-статистический метод семантической классификации допускает бóльшую формализацию и объективность анализа /по сравнению с другими методами/, он и был принят нами в качестве основного принципа построения динамического словаря условных синонимов.

В качестве лексического материала для обработки предлагается использовать термины и словосочетания поисковых предписаний информационно-поисковых систем /ИПС/, базирующихся на информационно-поисковом языке без лексического контроля, то есть без фиксированного словаря индексирования. Так как основной ИПС, обеспечивающей документальный тематический поиск в автоматизированной системе научно-технической информации /АСНТИ/ ЦНТИ в РСФСР, является ИПС РАСПРИ, то в качестве материала для описываемого словаря были взяты запросы этой системы /Общее количество запросов во всей АСНТИ составляет порядка 50 тысяч, каждый из которых включает в себя 20-30 поисковых терминов/. Запрос ИПС РАСПРИ представляет собой конъюнктивную нормальную форму, где каждый член конъюнкции, представляющий собой множество поисковых признаков /слов, словосочетаний, классификационных индексов и др. /, находящихся в отношении условной /поисковой/ синонимии, описывает один смысловой аспект запроса.

Разработаны алгоритмы и программы, которые на множестве лексических единиц сводного массива запросов ИПС РАСПРИ на машиночитаемых носителях определяют частоту совместной встречаемости отдельных лексических единиц в рамках какого-то одного смыслового аспекта в разных запросах. Далее путем анализа этой частотной характеристики из всего набора лексики формируется множество непересекающихся классов условных синонимов. Классом условных синонимов с критерием N в таком случае считается множество таких терминов из массива запросов, которые попарно втречаются совместно в рамках одного аспекта не менее, чем N раз.

Можно сказать, что данный механизм представляет собой своеобразную экспертную систему, делающую выводы на основе оценок разных экспертов /в нашем случае оценивается отношение синонимичности запросов, а "экспертами" выступают составители этих запросов/.

Сформированный таким образом массив лексических единиц фактически представляет собой словарь поисковых синонимов с указанием степени семантической близости отдельных лексем. Такой словарь может использоваться как для автоматического расширения запросов, так и в процессе диалога пользователя с ЭВМ. Для последнего случая разработана специальная программа, которая на введенные пользователем с клавиатуры терминала термин и критерий семантической связанности N выдает все термины массива запросов, которые встречаются с данным введенным термином в рамках одного аспекта не менее, чем N раз. Таким образом, данный алгоритм позволяет разбивать весь набор лексики объемом M на M пересекающихся классов условных синонимов.

В любом случае с помощью такого словаря может быть значительно повышен коэффициент полноты информационного поиска.

Наполнение полученных вышеизложенными способами классов условно синонимичных терминов не всегда тривиально с точки зрения обычного носителя языка и даже специалиста. Иными словами, полученные результаты могут представлять определенный интерес не только с точки зрения улучшения качества функционирования ИПС, но также и с точки зрения специалистов в области лексикологии и лингвистической семантики.

ЗНИКИНА Л.С.

ПРОБЛЕМЫ ГОМОМОРФИЗМА ЕЯ И ИЯ В ЗАДАЧАХ СИТУАЦИОННОГО УПРАВЛЕНИЯ ГОРНОГО ПРОИЗВОДСТВА

Современный этап развития компьютерных систем ОАСУ - САПР - уголь в горном производстве ориентирован на применение персональных компьютеров и вычислительных систем, в которых интерактивный ЕЯ-режим является основным, поскольку работу непосредственно с ПЭВМ по принятию решений ведут инженеры-технологи, т.е. пользователи-непрограммисты. Наряду с требованием простоты общения человека с компьютером повышается требование и на качество принятия решений, в частности, при ситуационном управлении [1].

В этой связи вопрос о соответствии и адекватности ЕЯ-ИЯ является весьма важным и требует дальнейших лингвистических исследований. При этом актуальную научную задачу представляет исследование гомоморфизма естественного (ЕЯ) и искусственного (ИЯ) языков как семантических систем, так как от степени адекватности ЕЯ-ИЯ существенно зависит "качество" лингвистической трансляции.

Однако с точки зрения практических задач, решаемых в системе "ЕЯ-ЭВМ", возникает проблема "допустимой неадекватности", обусловленный факторами семантической чувствительности и избирательности в системе "ЕЯ-ИЯ" [2].

Очевидно, оптимальному количеству допустимой потери семантической информации в системе "ЕЯ-ИЯ" соответствует оптимальная степень гомоморфизма G (ЕЯ/ИЯ).

Применительно к гомоморфизму ЕЯ/ИЯ можно рассмотреть две системы $J\langle\ \rangle$ и $J'\langle\ \rangle$, причем система J' (ИЯ) представляет гомоморфную лингвистическую модель системы J (ЕЯ).

Лингвистическая модель содержит информацию о морфологии (М), синтаксисе (С) и семантике (S) подмножества ЕЯ. Семантика определяется как интерпретация компонентов ЕЯ-текста компонентами модели проблемной среды.

Тогда система $J'\langle M', C', S'\rangle \subset M' \times C' \times S'$ называется моделью системы $J\langle M, C, S\rangle \subset M \times C \times S$ в том и только в том случае, когда

$$(\forall (M, C, S))\big((M, C, S) \in J \Rightarrow g(M, C, S) \in J'\big)$$

Здесь $g(M, C, S)$ - гомоморфизм ЕЯ/ИЯ:

$$g(M, C, S) = \big(g_m(M),\ g_c(C),\ g_s(S)\big),$$

где гомоморфизм $g_M(M), g_C(C), g_S(S)$ есть отображения:

$$g_M : M \to M'; g_C : C \to C'; g_S : S \to S'$$

В случае, если языки ЕЯ и ИЯ являются адекватными, выполняется условие:

$$(\forall(M', C', S'))((M', C', S') \in J \Rightarrow g^{-1}(M', C', S') \in J)$$

В докладе изложена концепция построения гомоморфной лингвистической модели системы общения "человек-ЕВМ" для ситуационного управления объектами горного производства. В соответствии с этим решаются следующие задачи:

- развитие семантической концепции гомоморфизма ЕЯ-ИЯ, реализующих диалог "человек-ЭВМ";
- установление критериев семантической чувствительности и полноты лингвистической модели ПО.

Результаты этих задач позволяют сформулировать принципы построения ЕЯ-диалог на основе исходной семантической информации и последующей ее трансляции в формализованную модель системы ситуационного управления объектами горного производства с помощью персональных компьютеров IBM PC - XT (AT). При этом качество функционирования системы определяется степенью совпадения смысла ЕЯ-предложений с их формализованным представлением, обеспеченным заданной степенью гомоморфизма.

Л И Т Е Р А Т У Р А

I. Поспелов Д.А. Ситуационное управление. Теория и практика. - М.: Наука, 1986. - 284 с.

2. Перспективы развития вычислительной техники. Под редакцией Ю.Н. Смирнова. Кн. I. Информационные семантические системы. Н.М. Соломатин. - М.: Высшая школа, 1989. - 127 с.

КААЛЕП Х.-Я.

ДЕМОНСТРАЦИОННЫЙ ОБРАЗЕЦ ЯЗЫКОВОГО ИНТЕРФЕЙСА ДЛЯ ПРИЗ-А

Программа переводит описанную на естественном языке текстовую задачу на язык УТОПИСТ, а задачу на языке УТОПИСТ, в свою очередь, решает ПРИЗ.

Программа не универсальна, а "понимает" тексты только определенной области знаний.

Текстовые задачи, из которых описываемая в настоящей работе программа стремится извлечь (вычитать) систему уравнений, относятся к крайне узкой области. Это кинематические задачи из курса школьной физики (7–8 классы школы с русским языком обучения), касающиеся равномерного движения.

Программа имеет модульную структуру и состоит из следующих частей:

1. Морфологический анализ.
2. Синтаксический анализ.
3. Семантический препроцессор.
4. Семантический анализ.
5. Формирование текстового файла на языке УТОПИСТ.

Морфологический и синтаксический анализ универсальны, т.е. они не предусмотрены для анализа текстов только в узкой предметной области. Они могут быть использованы как модули при построении других лингвистических процессоров или языковых интерфейсов. Остальные три части программы тесно связаны с предметной областью – с задачами по кинематике.

1. Морфологический анализатор ограничивается рассмотрением слов как структур вида "основа" + "флексия". Для омонимичных слов выдается альтернативный список основ с набором грамматической информации для каждой из них.

Морфологический анализ основан на словаре, который универсален и содержит 5000 наиболее часто употребляющихся в русском языке корневых слов.

2. Синтаксический анализ состоит из двух этапов:

1) Преобразование сложных предложений в простые.

2) Синтаксический анализ простых предложений (составление деревьев синтаксической зависимости).

Формирование простых предложений из сложных – полностью самостоятельный блок. Он основывается на правилах

препинания и грамматике слов. Анализ достаточно груб - получаемые простые предложения грамматически не всегда достаточно корректны.

На втором этапе синтаксический анализ опирается на выявленные в ходе морфологического анализа традиционные грамматические категории, используя при этом лишь некоторые из них. Из грамматической информации, выявленной в ходе морфологического анализа, часть остается в стороне.

Выход синтаксического анализа - деревя синтаксической зависимости, причем не отмечено, с зависимостью какого вида мы имеем дело в каждом конкретном случае.

3. Семантический препроцессор приводит деревья синтаксической зависимости к такому виду, чтобы в ходе семантического анализа они легче поддавались обработке.

4. Семантический анализ осуществляется следующим образом. Предложение за предложением анализируют синтаксические деревья. При этом делается "чертеж". "Чертеж" - это структура данных, имитирующая чертеж на бумаге, который может быть сделан на основе текстовой задачи. На "чертеже " описываются исходный и конечный пункты движений, двигающиеся, начальный и конечный моменты движения, расстояния между точками, время и скорость движения. На "чертеже" может быть описана, например, и средняя скорость, суммарное время на пути и прочие числовые величины, встречающиеся в тексте задачи.

Семантический анализ состоит из ряда подпрограмм, которые запускаются, если при прохождении синтакс. дерева встречается слово определенного типа. При каждом слове на синтаксическом дереве программа анализирует место слова на дереве (т.е. то, какие слова подчиняются ему и каким - подчиняется оно само), морфологический вид слова, его семантику и нарисованный к этому моменту "чертеж".

5. По "чертежу" необходимо извлечь (вычитать) переменные, скрытые в тексте данной задачи, и уравнения, необходимые для решения задачи. На основе "чертежа" и генерируется текстовый файл на языке УТОПИСТ, в котором содержатся уравнения и переменные, которым в тексте задачи присвоено некоторое значение (вместе с некоторыми комментариями относительно того, откуда эти переменные поступают).

Полученный текстовый файл на языке УТОПИСТ является вводом для ПРИЗА, на его основе ПРИЗ и решает задачи.

КАЗАКЕВИЧ О.А.

ТЕКСТОВАЯ БАЗА ДАННЫХ МЛАДОПИСЬМЕННОГО ЯЗЫКА: СОЗДАНИЕ, ИСПОЛЬЗОВАНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

Необходимость интенсифицировать изучение бесписьменных и младописьменных языков, с одной стороны, и возможности, предоставляемые для этого ЭВМ, с другой, привели к идее создания машинных фондов бесписьменных и младописьменных языков. В качестве модельного образца таких фондов в лаборатории АЛС НИВЦ МГУ построен машинный фонд селькупского языка.

Основными компонентами машинного фонда селькупского языка являются текстовая и словарная базы данных. Поскольку селькупская лексикография, как и лексикография большинства бесписьменных и младописьменных языков, развита слабо, на начальном этапе работы приоритетным направлением стало развитие текстовой базы данных.

Единственной формой существования бесписьменных языков и основной - младописьменных является устная речь. В связи с этим формирование текстовой базы целесообразно начинать с корпуса текстов устной речи, причем предпочтение следует отдавать неопубликованным материалам, чтобы обеспечить их скорейший ввод в научный обиход. В текстовую базу данных селькупского языка вошел корпус фольклорных текстов на тазовском диалекте селькупского языка объемом 10196 словоупотреблений и пофразовый перевод этих текстов на русский язык. Корпус состоит из 28 текстов, взятых из архива экспедиций к тазовским селькупам отделения структурной и прикладной лингвистики филологического факультета МГУ, проводившихся в 1970-1977 гг. Объем текста колеблется от 66 до 1624 словоупотреблений. Основная часть корпуса отражает среднетазовский говор, на который ориентирована селькупская письменность. Наряду с этим в корпусе представлен верхнетазовский говор. По тематике тексты распадаются на шесть групп: мифы, бытовые сказки, героические сказки, рассказы о великих шаманах, волшебные сказки, сказки о животных. Текстовая база данных реализована на ЭВМ ЕС 1036. В качестве программного обеспечения используется тексто-ориентированная компонента автоматизированной лексикографической системы УНИЛЕКС, разработанная Ж.Г.Мошкович [3].

Чтобы расширить возможности использования корпуса, была проведена его лемматизация (сведение встретившихся в текстах

словоформ к исходным словарным формам). Лемматизация велась автоматически с последующим исправлением ошибок в интерактивном режиме. При каждой словоформе корпуса указывается теперь ее словарная форма, часть речи, к которой она относится, и минимальная грамматическая информация (вид для глаголов, одушевленность и указание имен собственных для существительных и т.д.) [I].

На текстовой базе данных получены частотные словари и словоуказатели словоформ и лексем как по всему корпусу, так и по отдельным говорам и тематическим группам, а также конкордансы с переводом контекстов на русский язык. Проведен подсчет встречаемости в корпусе различных фонем и типов фонем.

Несмотря на то, что объем корпуса сравнительно невелик, результаты его машинной обработки уже нашли практическое применение. Полученные словари и конкордансы были использованы при подготовке грамматики селькупского языка [2]. На материале корпуса проведены исследования изобразительной лексики и способов выражения причинных и целевых отношений в селькупском тексте.

В настоящее время ведется работа по расширению текстовой базы данных. Текстовая база пополняется за счет корпуса фольклорных текстов баишенского говора тазовского диалекта селькупского языка, записанного Л.А.Варковицкой в 1941 г. Объем корпуса превышает 30 тыс. словоупотреблений. Баишенский корпус представляет собой необработанные полевые записи текстов, поэтому на первом этапе его машинной обработки предстоит нормализовать написание словоформ и только после этого перейти к лемматизации. С помощью ЭВМ предполагается подготовить корпус Л.А.Варковицкой к публикации.

Л И Т Е Р А Т У Р А

I. Казакевич О.А. Структура и информационное обеспечение машинного фонда селькупского языка // Информатика вычислительных систем. - М.: Изд-во Моск. ун-та, 1990. - С. 33-41.

2. Кузнецова А.И., Грушкина Е.В., Казакевич О.А., Хелимский Е.А. Учебник селькупского языка. Для педучилищ. - Л.: Просвещение (в печати).

3. Мошкович Ж.Г. Автоматизированная лексикографическая система УНИЛЕКС-2. - М.: Изд-во Моск. ун-та, 1989.

КАЗАКОВ М. И., ЯБЛОНСКИЙ С. А.

КОРРЕКТОР РУССКОГО ТЕКСТА РУССИКОН-I

Корректоры текста в настоящее время применяются практически во всех текстовых редакторах зарубежных фирм, таких как Word Microsoft, WordPerfect, Wordstar Professional, а также используются в виде отдельных интеллектуальных пакетов типа Grammatik III, IV и др., совместимых с тектовыми редакторами.

Корректор РУССИКОН-I позволяет:

- проверять текстовые ASCII-файлы, создаваемые любым текстовым редактором;
- выявлять и исправлять орфографические ошибки текста;
- выявлять и исправлять ошибки согласования слов в предложениях текста;
- создавать множество правильных слов-кандидатов на замену слова с орфографической ошибкой;
- создавать два типа пользовательских словарей, куда могут помещаться слова, отсутствующие в базовом словаре корректора: а) словарь имен, иностранных слов и аббревиатур; б) словарь терминов предметной области.

В состав корректора входят:

- базовый словарь русского языка и комплекс служебных словарей;
- морфологический анализатор русского текста;
- система многооконного интерфейса пользователя.

Скорость проверки текста определяется двумя основными факторами: техническими характеристиками персонального компьютера типа IBM PC и об'емом базового словаря корректора. Для IBM PC/AT с об'емом ОП 1-2 МБ и словарем об'емом 30 тыс. словоизменительных основ средняя скорость проверки составляет 20 слов/с.

Программа корректора написана на языке СИ и требует не более 64 К оперативной памяти.

КАПАНАДЗЕ О.Г.

ЭКСПЕРИМЕНТАЛЬНАЯ СИСТЕМА НЕМЕЦКО-ГРУЗИНСКОГО МАШИННОГО ПЕРЕВОДА

Многоуровневая модель машинного перевода, которая послужила теоретической базой построения системы, основывается на следующих предпосылках:

1.Стратификационная модель машинного перевода /МП/ должна опереться на многоуровневом анализе и синтезе, не ориентированном на конкретный входной и выходной естественные языки.

2.Процедуры анализа и синтеза должны быть реализованы на основе трансферного метода, предусматривающего одновременно возможность унифицированного представления каждого из уровней описания естественного языка /ЕЯ/.

3.Исходя из принципов, разработанных в данной стратификационной модели МП, унифицированное представление уровней описания ЕЯ обеспечивает использование любого из них не только для вертикального трансфера, но и в качестве интерфейсной структуры.

4.Интерфейсная структура должна предоставлять возможность как диагонального трансфера в многоуровневой модели МП с асимметрическим анализом и синтезом, так и горизонтального трансфера с неполным анализом и синтезом в рамках той же модели для ЕЯ с близкой структурой.

5.Построение стратификационной модели МП должно обеспечивать наращивание языковых пар на основе разработки и подключения только новых двуязычных интерфейсных структур, необходимых для горизонтального или диагонального трансфера, наряду с модульным построением процедур анализа и синтеза вновь подключаемых ЕЯ.

Предложенная многоуровневая модель МП отличается от других подходов к машинному переводу тем, что во время построения практических систем МП располагает большей мобильностью. Данное преимущество достигается с помощью модульного построения процесса анализа и синтеза, а также в результате разработки подмодулей,соответствующих описаниям отдельных уровней входного и выходного естественного языков. Это, в свою очередь, позволяет не ориентировать процедуру анализа входного языка на структуру и особенности выходного языка и наоборот.

В отличие от других моделей МП, основанных на использование языка-посредника /интерлингва/, которые зачастую предусматривают многоуровневый анализ и синтез ЕЯ, построенная в данном исследовании многоуровневая модель МП тесно связывается с трансферным

методом.

Благодаря введению жесткого требования унифицированного описания всех уровней ЕЯ, наряду с возможностью вертикального трансфера /с одного уровня входного или выходного языка на другой уровень того же языка/, предусматривается принципиальная возможность трансфера любого уровня описания входного ЕЯ в любой гомогенный /или толерантный/ уровень выходного ЕЯ. Соответственно, на основе понятий горизонтального и диагонального трансфера в рамках стратификационной модели МП с трансфером, вводится понятие модели МП с симметрическим и асимметрическим анализом и синтезом.

Использование рассмотренных принципов построения стратификационной модели МП повышает эффективность процесса создания реальных систем и их способность функционировать в мультиязыковом режиме.

На основе предложенных принципов реализован процессор грузинского языка многоцелевого назначения, для которого разработаны:

- уровень морфологического описания грузинского языка на основе теоретико-множественных моделей языков. Используя ранговый метод распределения морфем разработаны соответствующий алгоритм и программа на языке программирования PL/I;
- уровень поверхностно-синтаксического описания грузинского языка, основанного на принципах лексико-функциональной грамматики и обобщенной грамматики фразовой структуры;
- уровень глубинно-синтаксического представления, в основе которого положена лингвистическая теория валентностей Теньера;
- уровень семантической структуры на основе данных, полученных в процессе описания вышеперечисленных уровней, который рассматривается в качестве интерфейсной структуры при трансфере для экспериментальной системы МП грузинский язык-немецкий язык.

Программное обеспечение для последних трех уровней выполнено на языке программирования *C-Prolog*;

Идеи и результаты данного исследования нашли применение при реализации стратификационной модели МП во время совместной работы в ФРГ с Саарбрюкенской группой, занятой созданием модулей анализа и синтеза немецкого языка в рамках проекта машинного перевода ЕВРОТРА. В процессе работы над двуязычной интерфейсной структурой мы не ставили целью получение высоких показателей качества перевода, несмотря на то, что последующий анализ показал большие резервы повышения качества при углубленной работе над контекстуальными соответствиями лексических единиц.

КАРПИЛОВСКАЯ Е.А.

СТРОЕНИЕ И ФУНКЦИИ АВТОМАТИЧЕСКОГО ЧАСТОТНО-ВАЛЕНТНОГО СЛОВАРЯ МОРФЕМ

I. Проблемы отождествления алломорфов и размежевания морфов-омографов являются кардинальными для морфемики флективных языков. Создание строгих и эффективных процедур для решения указанных задач позволяет автоматизировать анализ морфемных структур слов, а также синтез единиц, морфемное строение которых допустимо системой данного языка. Подобные процедуры, как показал опыт компьютерной лексикографии, имеют первостепенное значение при составлении с помощью ЭВМ морфемных и словообразовательных словарей /3; 4/. Существует несколько подходов к формированию морфных рядов и определению их доминанты - морфемы-инварианта /I, с.35-36/.Общим в них является следующее требование к членам морфного ряда: их формальные отличия должны подчиняться действующим в данном языке морфонологическим правилам. Для составления таких правил, которые обладали бы достаточной описательной силой, необходима исчерпывающая информация о количественных, комбинаторных, позиционных и частотных характеристиках отдельных морфов. Сведения о дистрибуции и спектре частеречного распределения морфов позволяет установить формальные критерии для размежевания омографических единиц.

2. Автоматический частотно-валентный словарь морфем /далее - ЧВС/ содержит сведения о частоте, комбинаторике и частеречном распределении морфов различных классов. Он включает следующие самостоятельные подразделы: словари корней, префиксов, суффиксов, флексий и межкорневых прокладок. Данный словарь является продуктом обработки фактической базы машинного морфемно-словообразовательного фонда украинского языка /2/. Словник каждого из словарей-компонентов ЧВС на первом этапе исследований состоит из отдельных морфов. Единицы словников снабжаются показателями их продуктивности в исследуемом материале и удельного веса в нем. Для совокупности тех слов с данным морфом, которые представлены в частотном словаре украинского языка, указывается суммарная абсолютная частота их употребления в текстах. В информационный кортеж заглавной единицы статьи включается также перечень частей речи, в словах которых зафиксирован данный морф. Основная часть статьи ЧВС морфем состоит из пре-

позитивного и постпозитивного отделов. Первый содержит набор пар элементов в которых морф встречается в препозиции, а второй - пары элементов с этим же морфом в постпозиции. Каждая пара морфов сопровождается количественными, частотными и частеречными показателями, уточняющими информационный кортеж к заглавной единице статьи. В качестве иллюстраций приводятся простые и сложные слова тех частей речи, в которых обнаружены данные морфные пары. После того как наборы пре- и постпозитивных морфных пар сформированы, они автоматически упорядочиваются в статье по спаду количественных показателей. На следующем этапе исследований для формирования морфных рядов составляются специальные морфонологические правила, которые действуют с учетом представленных в ЧВС сведений о частоте и комбинаторике формально связанных морфов. Информация о "поведении" морфов в языковой системе служит также для разработки процедуры автоматического размежевания омографических единиц.

3. Автоматический частотно-валентный словарь морфем /ЧВС/ служит фактической базой и инструментом для решения не только рассмотренных актуальных проблем украинской морфемики, но и позволяет определять диагностирующие признаки морфемного строения слов отдельных частей речи, что имеет важное значение для автоматического синтеза слов. ЧВС используется также для создания автоматического морфемного сегментатора слов. Последний применяется в процессе автоматического пополнения фактической базы морфемно-словообразовательного фонда.

Л И Т Е Р А Т У Р А

I. В.Б. Касевич. Морфология. - Л.: Изд-во ЛГУ, 1986, 160 с.
2. Н.Ф. Клименко, Е.А. Карпиловская, Л.И. Комарова, Т.И. Недоизм, Т.В. Иванова. Морфемно-словообразовательный фонд украинского языка: принципы организации и перспективы использования // Третья Всесоюз. конф. по Маш. фонду рус. яз., М.: Ин-т рус. Яз. АН СССР, 1989, ч. П, с. 140-143.
3. З.Ф. Оливериус. Морфемы русского языка: частотный словарь. - Universita Karlova, Praha, 1975, 175 с.
4. Д.С. Уорт. Русский словообразовательный словарь: Введение // Новое в заруб. лингвистике, М.: Прогресс, 1983, вып. XIУ, с. 201-226.

КАРПОВ В.А.

ОКТЕТНАЯ УПАКОВКА СЕМАНТИКИ В БАЗЕ ЗНАНИЙ

В рамках Республиканской программы "Информатика" НИЛ теоретической и прикладной лингвистики БГУ им. В.И.Ленина разрабатывает концепцию человекомашинного интерфейса. Особенностью этой концепции является то, что ее разработка ведется на базе созданной общей теории языка, использующей учения и алгоритмы общей теории систем Ю.А.Урманцева. Согласно общей теории языка язык является симметрично-асимметричной системой, предполагающей изоморфизм и полиморфизм.

Использование свойств симметрии/асимметрии при создании баз знаний на естественном языке оборачивается возможностью построения иерархо-неиерархических систем с изоморфными структурными свойствами. При градуированном переходе от одного семантического класса к другому это позволяет решать проблему метафоризации. Так, если за классом I закреплено множество признаков I, а за классом 2 - множество признаков 2, то при анализе композиции с кодами 2 - 2, I - I фиксируется норма, а при кодах 2 - I и I - 2 методом вычитания обнаруживается метафоризация первого порядка. Далее - аналогично.

Важной особенностью стратегии упаковки семантики является тот факт, что мы ориентируемся на предельно максимальную упаковку знаний в коде слова, что дает возможность программе работать прямо с текстом без обращения к специальной базе знаний. Другими словами - наиболее общие знания о мире "зашиваются" в слове, при этом мы стремимся не выходить за типоразмер в 58 байт.

Выбор структур упаковки знаний прямо связан с предсказывающими возможностями этих структур, с их компактностью, с их изоморфизмом. Разработанная нами теория кубов чисел позволяет строить изопараметрические пространства неограниченного объема при задании трех векторов. При задании исходного и конечного вершинных чисел спецпрограмма выстраивает всю структуру элементов семантической сети, т.е. сеть разворачивается на базе имеющегося словаря, пустые вершины могут заполняться позже.

Исходное число отношений, которые можно задать на кубе чисел, равно I3, при необходимости это число может быть увеличено делением грани. Пространство из 27 кубов содержит внутри гиперкуб, внутри пространства из 729 гиперкубом будет являться пространство из 27 кубов чисел. Такой подход дает выход в топологию вложенных пространств, дифференциальную геометрию с их разработанным теоретическим аппаратом.

КЛИМЕНКО Н.Ф., КАРПИЛОВСКАЯ Е.А., КОМАРОВА Л.И.

МАШИННЫЕ СЛОВАРИ МОРФЕМНО-СЛОВООБРАЗОВАТЕЛЬНОГО ФОНДА УКРАИНСКОГО ЯЗЫКА

Морфемно-словообразовательный фонд формируется как составная часть Машинного фонда украинского языка, выполняющая функции информационно-справочной, исследовательской и редакционно-издательской системы. По своим исследовательским задачам он близок к международному проекту "Языковая синергетика" /1/. Описываемый фонд представляет в упорядоченном виде данные о морфемном и словообразовательном уровнях украинского языка, а также позволяет автоматически конструировать на их основе морфемные и словообразовательные словари разных типов. Кроме того, он выполняет роль Генерального словника украинского языка. Этот словник составляет интегральную часть словарно-текстового фонда украинского языка, его каталог. В настоящее время в машинной форме он содержит словник украинского языка по данным четырех словарей - толкового, морфемного, частотного, иностранных слов. В ближайшее время предполагается его расширение за счет орфографического словаря и "Словаря-справочника по правописанию и словоупотреблению" /2/.

По существу, речь идет о создании комплексного словаря украинского языка, в котором возле каждого слова формируется информационный кортеж, организованный по зонному принципу. Однотипные данные записываются в зоны с одинаковыми метками. Сейчас кортеж содержит сведения о наличии слова в одном из четырех введенных в ЭВМ словарей, а также данные об ударении, членении слова на морфемы, его частеречной принадлежности, количестве значений по толковому словарю, частоте в художественной прозе, количестве всех морфем и корней в частности. После подготовки грамматического словаря в особой зоне появится информация о типе парадигмы слова. Планируется формирование зон по этимологии слова, его возрасту. После создания текстового фонда украинского языка возможно подключение к этой информации документального фонда примеров использования слова в текстах.

Программно реализованный "Словарь символьных моделей слов украинского языка" позволил приписать каждой единице еще и формулу ее морфемного строения, которая записана в терминах классов морфем, входящих в состав слова. Так, расчлененное на морфемы слово до-гляд-а-ти перезаписано машиной как PKSS . Эта формула стано-

вится заглавной единицей упомянутого машинного словаря. Она сопровождается статьей, содержащей следующие сведения: абсолютная частота, удельный вес слов с этой формулой по данным словарей и текста, иллюстрации. Словарь этого типа создает предпосылки для изучения морфемики украинского языка как системы, законов строения слова. Определен инвентарь моделей морфемных структур слова. В пределах 100 тыс.единиц обнаружено 366 моделей. Из них 50 моделей свойственны простым словам /с одним корнем/, остальные - сложным. Стало возможным изучение этой системы с точки зрения наличия в ней ядра и периферии, симметрии//асимметрии. При построении морфемных структур определяющую роль играют законы глубины слова, простоты, предпочтения, симметрии /4/. По признаку "возможность развертывания препозитивной /докорневой/ и постпозитивной /послекорневой/ частей" структура простого слова украинского языка асимметрична. Подтверждается мысль В.Ингве об асимметрии структуры не только предложения, но и слова /3/. Возможности развертывания структур в препозитивной части намного уже тех, которые существуют при развертывании постпозитивных частей. Подавляющему большинству простых слов /74435, т.е. ≈ 91,7 %/ присуща асимметрия. Лишь 7,8 % из них обнаруживают по этому признаку зеркальную симметрию, при которой префиксальная часть слова как бы уравновешена суффиксально-флективной: РКF /за-сад-а/, РРК*SF* /до-о-чищ-ени-я/.Морфемная структура сложного слова строится иначе. Приблизительно 57,1% единиц этого типа построена по законам ритмичной, монотонной и в меньшей степени зеркальной симметрии.

Морфемно-словообразовательный фонд позволяет составлять и другие типы машинных словарей. Ведется работа над созданием "Частотно-валентного словаря морфем", подготовка которого завершается в 1991 г. Начата подготовка украинского корнеслова. На пробном массиве получены статьи гнездового корневого словаря с вариантами корня, членами гнезда, количественными показателями словесного наполнения.

ЛИТЕРАТУРА

1. Altmann G. und and.Projekt "Sprachliche Synergetik". -Bochum:Ruhr-Universität Verlag. 1985. - 27S.

2. Головащук С.І. Словник-довідник з правопису та слововживання. - Київ:Наук.думка, 1989. - 830 с.

3. Ингве В. Гипотеза глубины // НВЛ, М.:Прогресс, 1965, вып.IУ, с.126-138.

4. Перебийніс В.С. Кількісні та якісні характеристики системи фонем сучасної української літературної мови. -К.:Наук.думка,1970.

КОБРИН Р.Ю.

КОГНИТИВНОЕ МОДЕЛИРОВАНИЕ ПРЕДМЕТНЫХ ОБЛАСТЕЙ И ТЕРМИНОЛОГИЙ С ИСПОЛЬЗОВАНИЕМ ЭВМ : ОПЫТ МЕТОДОЛОГИЧЕСКОГО И ПРАГМАТИЧЕСКОГО АНАЛИЗА

Когнитивную модель можно определить как ориентированное на познание различных сторон объективной действительности формализованное знаковое представление знаний об объектах действительности, их свойствах, отношениях между объектами, включая действия объектов.

В основе когнитивного моделирования лежат знания. Понятие "знание" в современной информатике является развитием традиционного понятия "данные" с тремя качественными особенностями: наличием классификационных связей, ситуативных отношений и внутренней интерпретируемости [1]. Знания есть результат человеческого мышления и всегда основываются на системе понятий. Поэтому когнитивные модели формально представляют понятийные системы различных фрагментов объективной действительности.

Фрагменты объективной действительности /предметные области/ отражены в языке, специальные предметные области науки, техники, профессиональной деятельности - в терминологии, рассматриваемой как коммуникативная подсистема языка [2,с.42]. При переходе от действительности к языку часть актуальной информации утрачивается, и коммуникативные подсистемы языка, в том числе и терминологии, являются моделями предметных областей, приближенно передающими их /предметных областей/ онтологическую организацию.

Язык - орудие познания, поэтому терминологию можно назвать гносеологической моделью социально-культурных, профессионально-технических, научных и управленческих областей действительности. Эта гносеологическая модель является основной когнитивной моделью, обеспечивающей формализацию понятийных систем, отображающих и познающих предметные области действительности.

Особо отметим, что в качестве когнитивных моделей могут выступать и разнообразные алгебро-логические модели, формализующие понятийные системы, например, исчисление предикатов, системы продукций и др. [3].

В современной информатике и компьютерной лингвистике широко используются термины "концептуальная модель", "инфологическая модель" применительно к системам лингвистического обеспечения информационных систем различных типов. Под этими терминами понимается полное информационно-лингвистическое описание объектов, отношений между ними, действий над объектами [4,с.13]. Из определения видно, что

"концептуальная модель" есть когнитивная модель, ориентированная на лингвистическое представление предметной области в информационных системах.

Терминологии, как и описываемые ими предметные области, являются фрагментами действительности и, в свою очередь, могут стать и становятся объектами когнитивного моделирования.

Когнитивными моделями терминологий являются словари различных типов, используемые в коммуникации "человек-человек" и "человек--ЭВМ".

Когнитивное моделирование в задачах информатики и -прежде всего - искусственного интеллекта, основывающихся на режиме естественного языка, "предполагает и - более того -требует представления о том, что такое естественный человеческий язык и как он работает" [5,с.7I].

Поэтому когнитивное моделирование как метод когнитологии /инженерии знаний/ непосредственно связано с теоретической и компьютерной лингвистикой и использует разнообразные лингвистические методы и методики /сетевое моделирование лексики, метод непосредственно составляющих и др./.

В докладе рассматриваются принципы, технологии и реализация на ЭВМ СМ-4 когнитивных моделей предметных областей графической информации, представленных в естественном языке и искусственных знаковых системах [6].

На примере терминологии информатики показана реализация когнитивного подхода при построении фрагмента тезауруса на ЭВМ IBM PC/AT.

ЛИТЕРАТУРА

I - Поспелов Д.А. Вводные замечания // Представление знаний в человеко-машинных и робототехнических системах.-М.,I984.

2 - Кобрин Р.Ю. О месте терминологии в системе языка ":терминология как коммуникативная подсистема языка // Прикладная лингвистика и автоматический анализ текста. Материалы конференции.-Тарту: ТГУ, I988.

3 - Попов Э.В. Экспертные системы. Решение неформализованных задач в диалоге с ЭВМ.- М.:Наука,I987 /§3.4 "Модели представления знаний"/.

4 - Кобрин Р.Ю. Лингвистическое описание терминологии как база концептуального моделирования в информационных системах. Автореф. ... дис. докт. филол. наук.- Л., I989.

5 - Звегинцев В.А. Язык и знание // Вопросы философии. I982, №I.

6 - Васин Ю.Г., Кобрин Р.Ю. и др. База знаний лесоустройства и её реализация в автоматизированной картографической системе//Научно--техническая информация. Сер.2, /в печати/.

КОБРИН Р.Ю.
ПРИЛУЦКИЙ А.М.

ТЕКСТОВАЯ БАЗА ДАННЫХ ПАМЯТНИКОВ ДРЕВНЕРУССКОГО ЯЗЫКА НА ПЕРСОНАЛЬНОЙ ЭВМ IBM PC/AT

Актуальными задачами создания машинного фонда древнерусского языка являются:

- создание текстовой базы данных по спискам рукописей XI-XVII в.в., включая составление словоуказателя древнерусских текстов;
- создание машиночитаемой версии текстов рукописей [I, с.96].

Использование современных средств вычислительной техники при изучении памятников древнерусского языка позволяет получить детальную лингвистическую информацию о текстах рукописей, включая статистическую информацию, выявить особенности формирования грамматической нормы с учётом диалектных особенностей вариантов.

Несомненный интерес для исследователей представляет язык агиографических памятников провинциальных вариантов, поскольку в нём сочетаются языковая традиция и диалектные особенности. Наиболее значительное проникновение диалектных норм в язык памятников приходится на XVII в., период активного формирования старорусского языка.

Едва ли не идеальным объектом для исследования взаимодействия диалектных норм и языка памятников представляется "Житие Макария Желтоводского". Протограф краткой /исходной/ редакции "Жития" восходит к XVI веку, пространная редакция "Жития" создаётся на основе краткой не раньше 1615 г. [2, с.291-293].

Для компьютерной обработки текста используется информационная система CART PLUS типа электронная картотека. Система реализована на ЭВМ IBM PC/AT. На базе системы создано оригинальное программное обеспечение для многоаспектной статистической обработки информации.

Для ввода информации разработаны модели для каждой части речи, предусматривающие ввод встретившегося в тексте слова с его грамматическими характеристиками. Определён формат карточек, выдаваемых на терминал и на печатающее устройство. Ввод текста с клавиатуры осуществляется специально разработанным кодом, в котором к буквам современного русского алфавита добавлены знаки для обозначения ударений, титлов, лигатур и т.д., а также специфических букв древнерусского алфавита. Ввод текста лист в лист и получение точной копии текста рукописи на машинном носителе [I, с.97] не предусмотрены.

На основе системы создана база данных, обеспечивающая хранение, поиск и выборку информации по различным ключам, определён формат карточек, выдаваемых на терминал и на печатающее устройство.

Программное обеспечение позволяет производить - в частности - следующие операции:

- выборку из базы данных слов, характеризующихся определёнными морфологическими и другими лингвистическими признаками;
- определение распространённости отдельных грамматических форм / например, определение числа существительных различных падежей, диалектных вариантов и т.п./;
- автоматическое составление частотного словаря и словников;
- автоматическое получение конкорданса сочетаемости слов и форм, а также другие исследования текста.

В процессе эксплуатации системы может проводиться выборка как по одному слову, так и по нескольким, что позволяет получать словари словосочетаний. Информационная система позволяет организовать неограниченное количество баз данных с объёмом информации в каждой до 65 500 карточек.

Фрагменты текстовой базы данных демонстрируются на ЭВМ IBM PC/AT.

Л И Т Е Р А Т У Р А

I - Горина Н.Л. Создание компьютерных версий древнерусских рукописей // Третья Всесоюзная конференция по созданию машинного фонда русского языка. Тезисы докладов / ч.I /.- М.: ИРЯ АН СССР, 1989.

2 - Пронырко Н.В. Житие Макария Желтоводского и Унженского // Словарь книжников и книжности Древней Руси второй половины XIУ-XУI в.в. - Л., 1988, ч.I .

КОЙТ М.Э.

О ВЕДЕНИИ ВОПРОСНО-ОТВЕТНОГО ДИАЛОГА С ЭВМ ПО ДАННОМУ ТЕКСТУ ЕЯ

0. В процессе обучения человека к иностранному языку учитель обычно, время от времени, проверяет, понял ли обучаемый прочитанный текст, в том числе, умеет ли он отвечать на заданные по тексту вопросы или, наоборот, способен ли он сам задавать такие вопросы. В роли обучаемого к естественному языку может выступать, в частности, и ЭВМ.

Рассматриваем диалог человека и ЭВМ по заданному тексту ЕЯ. Чтобы являться равноправным партнером для человека в таком диалоге, ЭВМ должна уметь решать следующие подзадачи:

- понять текст,
- понять вопросы, заданные человеком,
- найти из текста ответы на вопросы и отвечать человеку на ЕЯ,
- сама задавать вопросы на ЕЯ и оценить правильность ответов человека на них.

В настоящем докладе дается короткий обзор о решении этих проблем в диалоговой системе, разрабатываемой в Тартуском университете.

1. Понимание текста ЕЯ (в данном случае, эстонского), в частности, вопроса или ответа, со стороны ЭВМ означает построение его семантического представления. Семантическое представление предложения - это фрейм предложения, т.е. граф, вершинам которого соответствуют названия объектов и понятий (слова), входящие в предложение, а граням - семантические связи между ними. Семантическое представление текста, состоящего из нескольких предложений, т.н. фрейм текста - это сеть фреймов предложений.

ЭВМ выполняет морфологический, синтаксический и семантический анализ текста, выходом чего является фрейм текста.

Вопросы, задаваемые человеком к ЭВМ, бывают трех типов:

- общие вопросы ("Работает ли Ян на фабрике?"; ответ, в зависимости от текста: "да", "нет", "нет, в магазине" или "неизвестно");
- альтернативные вопросы ("Работает Ян на фабрике или в магазине?"; возможные ответы: "на фабрике", "в магазине", "нет", "нет, в театре" или "неизвестно");

- специальные вопросы ("Где работает Яан?", "Кто работает на заводе?").

2. Если ЭВМ выступает в роли отвечающего, то человек вводит свой вопрос, а ЭВМ анализирует вопрос и определяет его тип.

В случае общего вопроса ЭВМ во фрейме текста ищет (под)фрейм предложения, входящий во фрейм вопроса. Если такой подфрейм найдется, то ответом будет "да". Если найдется подфрейм, в котором некоторые, но не все заполнители слотов (т.е. метки вершин) совпадают с заполнителями соответствующих слотов фрейма вопроса, то ответом будет либо просто "нет", либо "нет" + предложение, синтезированное из найденного подфрейма. В противном случае ответ: "неизвестно".

Альтернативный вопрос ЭВМ заменяет на общие вопросы, количество которых равняется количеству альтернатив, и попытается отвечать на них вышеописанным образом. Ответом будет альтернатива, оказавшаяся истинной.

Отвечая на специальный вопрос (с вопросительным словом "кто", "когда" и т.п.), ЭВМ во фрейме текста ищет подфрейм, где заполнители всех слотов, кроме запрашиваемого, совпадают с заполнителями соответствующих слотов фрейма вопроса. Ответом в случае успеха является предложение, синтезированное из найденного подфрейма, или просто заполнитель запрашиваемого слота, а в случае неудачи - "неизвестно".

3. Если ЭВМ выступает в роли спрашивающего, то она из фрейма текста выделяет фреймы предложений, сделает из каждого столько копий, сколько в этом фрейме имеется слотов, а затем в каждой копии по одному заполнителю слота заменяет его на соответствующее вопросительное слово. Из полученных фреймов потом синтезируются предложения ЕЯ - специальные вопросы, которые ЭВМ будет задавать человеку по одному. Получив (однословный) ответ человека, ЭВМ определяет, совпадает ли это слово с заполнителем соответствующего слота во фрейме текста, и выдает человеку свою оценку: "правильно" или "неправильно".

4. К настоящему моменту на ЭВМ из вышеописанного реализован этап работы с фреймами, т.е. общение человека и ЭВМ пока возможна лишь на "языке фреймов". Соответствующие программы составлены на языке Turbo Pascal для IBM PC.

КОКИНА С.А.

РИЖСКИЙ ПРОЕКТ ДАЙН

Для углубленного изучения содержательной стороны народно-поэтических текстов - латышских народных песен (дайн) на современном уровне необходим скрупулезный анализ каждого значимого элемента. Чтобы обеспечить анализ текстов большого объема необходимо использовать ЭВМ. С 1986 года в Институте языка и литературы им.А.Упита Латвийской АН и на физико-математическом факультете Латвийского университета начата работа по созданию банка данных латышских народных песен.

Работы в этом направлении были начаты в 1965 году канадскими учеными - И.Фрейбергсом, В.Берзиньшем, К.Конраде, В.Вике-Фрейбергой. Составлен массив дайн Бостон-Монреаль, содержащий около 71 000 дайн Копенгагенского издания дайн.

В основу Рижского банка данных положен свод латышских народных песен - "Латвью дайнас" Кришьяниса Барона. Он содержит 217 996 текстовых единиц, т.е., четверостиший, что составляет приблизительно 4 360 000 словоупотреблений.

За три года работы над проектом разработаны принципы кодирования текстов, занесены в память ЭВМ (персональный компьютер класса PC) три из шести томов свода "Латвью дайнас" и указатель субстантивов, начата работа по созданию комплекса программ для исследования. Указатель субстантивов - это регистр всех использованных форм имен существительных с указанием номеров песен, в которых они были употреблены. С помощью указателя возможно сформировать циклы песен определенной тематики. Первые результаты, полученные при помощи ЭВМ дают возможность исследовать некоторые ранее малоизученные аспекты этого типа текстов на столь представительной выборке.

При дальнейшей работе намечается реализовать возможности воссоздания отдельных коллекций, распределения текстов и лексики по этнографическим (административным) регионам, создание полного указателя отдельных мотивов, эпитетов.

КОКОЧАШВИЛИ Т.Г.

О СВЯЗИ НЕКОТОРЫХ РЕЗУЛЬТАТОВ СРАВНЕНИЯ ЭМПИРИЧЕСКИХ РАСПРЕДЕЛЕНИЙ СЛОВОУПОТРЕБЛЕНИЙ В ТЕКСТАХ С ПСИХОФИЗИОЛОГИЧЕСКИМИ ПРОЦЕССАМИ

Применяя аппарат теории вероятностей к проблеме выявления эмпирических законов распределений словоупотреблений в тексте мы получили результаты, заключающиеся в том, что многие словоформы, взятые из ЧС, составленных на базе рассмотренных нами грузинских и английских массивов одновременно подчиняются нескольким теоретическим законам распределений [1, с.115].

С математической точки зрения этот факт можно объяснить наличием так называемых общих зон действия, этих теоретических законов распределений, т.е. статистические параметры общих зон действия этих законов относятся к каждому из рассматриваемых законов.

Однако, объяснение этому факту можно найти, учитывая нейрофизиологические и психологические процессы, происходящие в мозгу автора. В работе [2, с.52] доказывается, что если два разных образа памяти активируются всегда вместе, то с течением времени такие образы сливаются вместе. Только те образы, частоты активаций которых не коррелированы, имеют возможность длительного независимого существования, это рассуждение относится и к образам слов речи.

Предположим теперь, что в каком-то тексте идет речь о внутренней энергии. Автор постоянно употребляет фразу "формула внутренней энергии". "Внутренней" - прилагательное, его эмпирическое распределение совпадает с одним теоретическим законом распределения. "Энергии" - существительное, его эмпирическое распределение совпадает одновременно с несколькими теоретическими законами распределений. Однако, автор, подразумевая, что речь идет о "формуле внутренней энергии" может писать "формула энергии". Тогда эмпирический закон "энергии" и будет совпадать одновременно с несколькими теоретическими законами. Дело в том, что "внутренней" и "энергии" объединяются в один образ памяти, а тем не менее каждое это слово, употребленное в отдельности, подчиняется разным законам распределений. Данные рассуждения были проверены на грузинском массиве и нашли свое подтверждение.

Л И Т Е Р А Т У Р А

1. Кокочашвили Т.Г., Цилосани Т.П., Беришвили Г.Ш. - Результаты сравнения эмпирических законов распределений частей речи в грузинских и английских научно-технических текстах с пятью теоретическими законами распределений. Сб. Квантитативные аспекты системной организации текста, Материалы межвузовского семинара, Тбилиси, 1987.

2. Лебедев А.М. - Циклические коды памяти. Когнитативная психология. Материалы финско-советского симпозиума. М., Наука, 1986.

КОЛОДЯЖНАЯ Л.И.
ДОРОХИНА Л.В.

АВТОМАТИЧЕСКИЕ ОПЕРАЦИИ НАД ДВУЯЗЫЧНЫМ СЛОВАРЕМ

В сообщении обсуждается компьютерная версия эстонско-русского базового словаря [1], реализованная на персональном компьютере IBM PC/XT с целью отработки в общем виде типичных для двуязычных словарей операций, в частности операции создания словаря-перевертыша.

Для создания и исследования машинной базы двуязычного словаря использовался пакет программ "Автоматический филологический словарь", реализованный на языке FOXBASE+. Эта система предназначена для работы в автоматизированном режиме с филологическими словарями различного типа (толковыми, переводными, синонимов и т.п.). Система оперирует со словарем, опираясь на формальное описание словарной статьи. Система позволяет лексикографу описать структуру нового машинного словаря в элементарных и составных компонентах, указывая связи между ними, производить ввод статей в базу, редактировать статьи в базе, а также работать со словарем в диалоговом и пакетном режимах. Программа диалога обеспечивает "вход" в словарь по любой компоненте, об"явленной как ключевая, и просмотр заданных фрагментов статей. Программы пакетного режима обеспечивают проведение четырех базовых операций - ВЫБОРКИ, ПРОЕКЦИИ, ИНВЕРСИИ ПРОСТОЙ(обобщение операции создания словаря-перевертыша), ИНВЕРСИИ ТАБЛИЧНОЙ.

В машинной базе эстонско-русского словаря хранится более 700 исходных статей (статьи вводились по книге). Для машинной версии книжного словаря получены с помощью базовых операций - русско-эстонский словарь, списки слов по частям речи(на русском и эстонском языках), таблицы распределения слов по некоторым параметрам.

1. Базовый словарь эстонского языка. - Таллинн, Фирма "Коолиус". Институт языка и литературы АН ЭССР, 1989.

КОЛОДЯЖНАЯ Л.И.
ПОЛИКАРПОВ А.А.

КОМПЬЮТЕРНАЯ ВЕРСИЯ СИНОНИМИЧЕСКОГО СЛОВАРЯ

В сообщении обсуждается компьютерная версия "Вебстеровского нового словаря синонимов" - WNDS /I/, реализованная на персональном компьютере IBM PC /XT . Для создания и исследования машинной базы синонимического словаря использовался пакет программ "Автоматический филологический словарь", реализованный в среде FOXBASE+. Система "Автоматический филологический словарь" предназначена для лексикографов, производящих операции над словарем с помощью компьютера. Система оперирует со словарем, опираясь на формальное описание словарной статьи /2/. Программное обеспечение "Автоматический филологический словарь" позволяет лексикографу описать структуру нового машинного словаря в элементарных и составных компонентах, указывая связи между ними (аналог формальной грамматики), производить ввод статей в базу, редактировать статьи в базе, а также работать со словарем в диалоговом и пакетном режимах. Программа диалога обеспечивает "вход" в словарь по любой компоненте, объявленной как ключевая и просмотр заданных фрагментов статей. Программы пакетного режима обеспечивают проведение базовых операций -ВЫБОРКИ, ПРОЕКЦИИ, ИНВЕРСИИ.

При создании базы WNDS было использовано описание структуры его статьи, разработанное авторами ранее /2/. В машинной базе хранится около 2000 основных статей (для которых заглавное слово -доминантное в группе). Остальные 5000 статей генерируются автоматически при организации "входа" в словарь от компоненты "синоним". Машинная версия WNDS используется для исследования различных словарных параметров.

I. Webster's new dictionary of synonyms.-Springfield (Mass): Merriam, 1973.
2. Колодяжная Л.И. Автоматизированная лексикографическая система УНИЛЕКС. -М.: Изд-во Моск. ун-та., 1987.

КОЛОДЯЖНАЯ Л.И.
ПОЛИКАРПОВ А.А.
ШУМАРИНА И.В.

ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИИ СИНТАКСИЧЕСКОГО СЛОВАРЯ ДЛЯ ИССЛЕДОВАНИЯ СЕМАНТИЧЕСКИХ ХАРАКТЕРИСТИК ЛЕКСЕМ

"Синтаксический словарь" /репертуар элементарных единиц русского синтаксиса/" [1] - первый в лингвистике опыт представления синтаксического строя языка в словарных параметрах. Компьютерная версия "Синтаксического словаря " реализована на персональном компьютере класса IBM PC/XT в среде FOXBASE+ . Машинная база словаря состоит из двух основных файлов / атрибуты синтаксем и манифестации синтаксем/ и нескольких служебных /список предложно-падежных форм, список опорных лексем, список авторов примеров/. Программное обеспечение словаря позволяет оперировать с информацией, организуя различные "входы" в базу - вход от предложно-падежной формы, вход от значения синтаксемы, вход от опрной лексемы, вход от автора примера. Пользователь имеет возможность вести работу с базой в диалоговом режиме, просматривая информацию на экране, или - распечатывать информацию о синтаксемах, прегруппированную относительно заданного входа в словарь.

Информация, содержащаяся в словаре, используется для автоматизации филологических исследований в различных направлениях. Одно из них - исследование семантических характеристик классов лексем. С этой целью в автоматизированном режиме создается словарь лексем, входящих в состав иллюстрационных словосочетаний к синтаксемам /опорные лексемы/. Словарь опорных лексем, в котором в качестве атрибутов лексемы дается значение синтаксемы с указанием функции и позиции в предложении, будет использован для выявления корреляции между классами лексем и значениями синтаксем.

Л И Т Е Р А Т У Р А

1. Золотова Г.А. Синтаксический словарь: репертуар элементарных единиц русского синтаксиса. М.: Наука, 1988.

КОРОЛЕВ Э.И.

ГРАММАТИЧЕСКИЕ СРЕДСТВА СОВРЕМЕННЫХ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОРОВ

В настоящее время широкое распространение получили текстовые лингвистические процессоры (ЛП) самого разнообразного назначения: системы машинного перевода (СМП), информационного поиска (АИС и СУБД), коррекции текстовых искажений, естественно-языковые интерфейсы для анализа запросов и синтеза ответов в экспертных системах, АИС и СУБД, словарные базы данных (переводные, терминологические) и т.д.

Все эти виды ЛП используют различные средства автоматического анализа (морфологического (МА), синтаксического (СА), семантического (СемА)) и синтеза (семантического (СемС), синтаксического (СС), морфологического (МС)) предложений естественного языка, что приводит к выделению 64 классов грамматических средств современных текстовых ЛП.

Эти классы объединяются в четыре типа ЛП.

ЛП первого типа не используют указанных грамматических средств анализа и синтеза и включают довольно распространенные АИС "без грамматики" или с элементами грамматики: АСПИД, АСИНИТ, ИНИС.

ЛП второго типа используют только системы автоматического анализа: МА ("Скобки"), МА и СА (АРТ), МА,СА,СемА (КАСКАД, PSS).

ЛП третьего типа -- это синтезаторы текста из семантической сети: СМП МЕТЕО (СС и МС), модели Симмонса и Клейна (СемС и СС).

ЛП четвертого типа -- это системы автоматического анализа и синтеза предложений ЕЯ: СМП пословно-пообротного типа СИСТРАН, СИЛОД (МА,МС); СМП с развитыми процедурами морфологического и синтаксического анализа и синтеза GETA-ARIANE, Mu, ATLAS I, HICATS, PENSEE, TITUS IV, АСПЕРА ; системы "синтактико-семантического" машинного перевода ASCOP, SEMSYN, ROSETTA, DLT, LUTE, TAURUS, PIVOT , модели интеллектуального диалога с СУБД на ЕЯ(ПОЭТ), системы построения и ведения баз знаний (МА, СА, СемА, СемС, СС, МС).

КОРОЛЕВ Э.И., ПОЛИКАРПОВ А.А., ШЕРШОВА А.В.

АВТОМАТИЗИРОВАННАЯ СИСТЕМА ВЕДЕНИЯ СЛОВАРЯ ДЛЯ ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА

1.В основу словаря для лингвистического процессора АСПЕРА положены следующие принципы: а/единицей описания является лексико-семантический вариант /ЛСВ/; б/ЛСВ подвергается глубокой обработке с точки зрения его лексико-семантических и семантико-синтаксических характеристик. В качестве основных параметров, определяющих ЛСВ, использовались: а/представление основы слова; б/общие морфологические характеристики; в/семантические признаки; г/семантико-синтаксические характеристики /модель управления с многоуровневыми идентификаторами: имя семантической валентности, морфологическое выражение актанта, семантические признаки актанта, обязательность/факультативность актанта/; д/лексические функции. Хотя словарь создавался для специальной предметной области, набор параметров, определяющих ЛСВ, и их содержательная интерпретация имеют универсальный характер и могут быть использованы в различных прикладных системах автоматической обработки текста, а также для исследовательских целей.

2.Предмашинная обработка материалов заключается в составлении лингвистом словарной статьи в соответствии с выделенными параметрами с использованием анкетного шаблона и инструктивных материалов, разъясняющих принципы описания ЛСВ. Ниже приводится фрагмент словаря.

```
1 N1850
2 СЛОВО  адаптация
2а N ЛСВ 1
2б ОСНОВА адаптаци
3 МОРФ. СВЕДЕНИЯ С=ж=неод
4 СЕМАНТИЧЕСКИЕ ПРИЗНАКИ изме
5 МОДЕЛЬ УПРАВЛЕНИЯ
  ВАЛЕНТНОСТЬ N 1
5.1 СЕМАНТИЧЕСКИЙ ПАДЕЖ Obj
```

5.2 МОРФ. ВЫРАЖЕНИЕ АКТАНТА Срод
5.3 СЕМАНТИЧЕСКИЕ ПРИЗНАКИ АКТАНТА устр=лицо
5.4 ОБЯЗАТЕЛЬНОСТЬ /ФАКУЛЬТАТИВНОСТЬ/ фак

3.С целью формирования и ведения базы автоматического словаря был разработан программный комплекс на основе СУБД РЕБУС, реализованный на ПЭВМ IBM PC/XT(AT). Поля БД соответствуют структуре словарной статьи. Ввод информации осуществляется пользователем в процессе диалога с системой. Введенная информация подвергается автоматическому контролю по ряду параметров /проверка на дубли номеров записей, полноту словарной статьи, правильность ввода отдельных формализованных параметров /морфологической информации, в частности, сочетаемости предлогов с конкретными падежами , семантических признаков, наименований валентностей/ Словарная БД, а также результаты контроля могут быть распечатаны как в базовом формате, так и в виде текстового файла.

После редактирования и коррекции словарная БД может автоматически перекодироваться с целью непосредственного использования в лингвистическом процессоре. В программе перекодировки, позволяющей значительно ускорить работу и повысить качество по сравнению с процедурами ручного кодирования, предусмотрены морфологические варианты использования основы ЛСВ, в т.ч. для случаев чередования основ.

4.Разработанная система позволила подготовить 1000 словарных статей. Автоматизация подготовки словаря позволяет резко повысить скорость прохождения всей технологической цепочки и качество словарного материала за счет автоматизации контроля и перекодировки параметров словаря.

КОССЕК Н.В.

ИСПОЛЬЗОВАНИЕ ТЕКСТОВЫХ БАЗ ДАННЫХ В ТЕКСТОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ СЛАВЯНСКОГО ЕВАНГЕЛИЯ

Появление компьютерной лингвистики оказало существенное влияние на все аспекты языковедческих исследований, а также их различных приложений. В частности, сложная проблема текстологических отношений между историческими памятниками оказывается ближе к своему решению. Текстология центрального славянского памятника – Евангелия – является почти не исследованной, несмотря на многочисленные работы в этой области, проведённые в конце XIX – начале XX века у нас и за рубежом. Значительным шагом вперёд в этом направлении служит применение текстологического метода Э.Колвелла [3], при котором тексты памятников сопоставляются непосредственно друг с другом без помощи эталона. На первом этапе работы исследователем выделяются узлы разночтений среди описываемых памятников, на следующем – наименования, узлы, а также отдельные чтения кодируются цифрами. Эти цифровые данные обрабатываются на ЭВМ, причём устанавливается процент общих чтений между каждой парой памятников (или рукописей) от количества всех узлов разночтений. В нашем материале – 8 среднеболгарских Евангелий на фоне древнейших старославянских текстов – было выделено 150 узлов разночтений. Затем на компьютере автоматически проводится группировка памятников по степени их близости между собой, что наглядно представляется в таблице. Так, данные, полученные на ЭВМ по текстам евангелий от Матфея, Марка и Луки, своеобразно связаны с текстологическими группировками, которые были ранее получены на материале евангелия от Иоанна, ст. XI.1-45 [2]. С помощью введённых текстовых баз данных можно анализировать текстологические особенности сколь угодно большого количества евангельских текстов, что принципиально приближает решение проблемы текстологии славянского Евангелия как специфического памятника с контролируемой традицией [1, с.94].

ЛИТЕРАТУРА

1. Алексеев А.А. Проект текстологического исследования кирилло-мефодиевского перевода Евангелия // Советское славяноведение. 1985, №1

2. Коссек Н.В. К вопросу о кирилло-мефодиевском переводе славянского Евангелия // Вопросы языкознания. 1988, №2

3. E.Colwell. Studies in Methodology in Textual Criticism of the New Testament. Leiden, 1969

КОЦУБА Н.Я.

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ГРАММАТИЧЕСКИХ ХАРАКТЕРИСТИК НЕМЕЦКОГО СКАЗУЕМОГО

На первом этапе работы рассматриваемый алгоритм определяет, к какому типу относится анализируемое предложение, поскольку тип предложения определяет местоположение сказуемого. Это достигается путем поиска запятых и анализа их лексического окружения на наличие союзов или союзных слов, характерных для сложных предложений. Для случаев, когда придаточное предложение стоит перед главным, первое слово в фрагменте перед запятой проверяется на идентификацию с подчинительным союзом.

Данный блок алгоритма мы реализовали в программе INFO.BAS. Программа реализована на микро-ЭВМ ДВК-2М, написана на языке BASIC. Программа выполняет сортировку вводимых с клавиатуры предложений на простые и сложные и ведет статистические подсчеты.

На втором этапе работы осуществляется непосредственный поиск и выделение сказуемого. Для поиска изменяемой части сказуемого используется автоматический словарь глагольных форм. Поиск неизменяемой части сказуемого производится на основе морфологического анализа. Кроме того особенности расположения данной части сказуемого в предложении предоставляют дополнительные возможности для ее поиска и выделения.

На основании данного блока алгоритма нами составлена программа DIPL.BAS на языке BASIC, включающая 260 строк. Программа реализована на микро-ЭВМ ДВК-2М.

Действует программа следующим образом. Из файла на гибком магнитном диске (ГМД) считывается очередное предложение текста. Программа поочередно сравнивает все словоформы анализируемого предложения с глагольными формами из автоматического словаря, хранящегося в памяти ЭВМ. Для данного словаря были отобраны наиболее употребимые в немецких научно-технических текстах формы 15-ти глаголов. Каждая из них имеет в словаре код вида X000. Здесь первый символ (буква) обозначает принадлежность данной формы к списку форм определенного глагола. Второй символ (цифра) кодирует номер данной глагольной формы в списке форм данного глагола. Третья цифра кодирует лицо, четвертая - число данной глагольной формы.

В докладе детально рассматриваются другие особенности созданных алгоритма и программы и результаты реализации программы.

КРАУКЛИС Л.Г., КУКУШКИНА О.В., ПОЛИКАРПОВ А.А.

ПРОЕКТ МАШИННОГО ФРАНЦУЗСКО-РУССКОГО И РУССКО-ФРАНЦУЗСКОГО СЛОВАРЯ "ЛОЖНЫХ ДРУЗЕЙ ПЕРЕВОДЧИКА"

"Ложными друзьями переводчика" принято называть слова разных языков, сходные по форме, но различающиеся по значению. Эти слова представляют значительную трудность при изучении иностранного языка и при переводе с одного языка на другой, поскольку близость формы вызывает ложное отождествление по значению, причем непроизвольная тенденция к смешению этих слов настолько велика, что ошибки встречаются не только у начинающих изучать язык, но и у профессиональных переводчиков. Поэтому эта часть лексики требует особенного внимания.

В настоящее время формируется база сопоставительных данных для создания машинной версии французско-русского и русско-французского словаря "ложных друзей переводчика". Отбор лексики производится на основе средних толковых словарей: "Словаря русского языка, т.т.I-4", М., I985-I988 и французского словаря *Petit Robert*, P., I98I. Слова берутся во всем объеме значений, с толкованиями, примерами и переводами на другой язык, как это было сделано в "Англо-русском и русско-английском словаре "ложных друзей переводчика" /М., I969/ и в "Немецко-русском и русско-немецком словаре "ложных друзей переводчика" К.Г.М.Готлиба /М., I972/. Однако наш словарь базируется на более полном сопоставлении слов по большему числу параметров.

Машинная версия словаря позволяет при его подготовке более четко систематизировать материал, оперативно его пополнять, использовать как справочную базу в практике преподавания и изучения французского и русского языков русскими и французами и в теоретико-сопоставительном исследовании французского и русского языков.

При создании машинной версии словаря предполагается для каждой лексической единицы учитывать следующие параметры.

I. Часть речи.

2. Лексико-грамматические параметры.

Для существительных:

а/ род;

б/ *pl.v. tant. / sing. tant.*;

в/ одушевленность/неодушевленность;

г/ счетность/несчетность;

д/ конкретность/абстрактность;

е/ собирательность/единичность.

Для глаголов:

а/ вид;

б/ переходность/непереходность;

в/ требует/не требует прямого/косвенного дополнения;

г/ управление /предлог падеж/;

д/ недостаточность парадигмы;

е/ предельность/непредельность;

ж/ однократность/многократность;

з/ каузативность/некаузативность.

Для прилагательных:

а/ качественность/относительность;

б/ особенности образования степеней сравнения.

Для местоимений:

а/ лексико-грамматическая группа /личные, притяжательные и т.д./.

3. Стилистические различия.

4. Орфография /например: *groupe* /группа – одиночная/сдвоенная согласная/.

5. Словообразование /например: *architecte* /архитектор – отсутствие/наличие суффикса у слов с одним корнем/.

6. Сочетаемость /есть ли ограничения и какие/.

7. Семантика.

Для слова в целом:

а/ расхождение во всем объеме значений/расхождение в некоторых значениях.

Для отдельных значений:

а/ расхождение касается значения/оттенка значения;

б/ значения разные, но семантически близкие/значения не связаны семантически.

8. Толкование.

9. Эквивалент:

а/ точный;

б/ неточный;

в/ отсутствует.

IO. Наиболее типичные примеры употребления слов с переводом на другой язык.

Словарь реализуется на ПЭВМ IBM PC/XT в среде DBASE 3+.

КРИТСКАЯ В.И.

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ СВЯЗЕЙ И ГРАНИЦ ОБОСОБЛЕННЫХ КОНСТРУКЦИЙ

Предлагаемое исследование является частью работы по созданию системы АСА /Институт языковедения АН УССР/.

Обособленные единицы встречаются практически в каждом простом предложении и в каждой предикативной части сложных предложений. Наиболее распространены причастные обороты, они составляют свыше 90% всех обособленных единиц в исследуемых текстах; среди других - поясняющие, адъективные, деепричастные, сравнительные, обстоятельственные обороты и приложения.

Обособление синтаксических конструкций - это смысловое выделение их относительно контекста /в пределах предложения/. Тем самым устанавливается несколько смысловых уровней в предложении: основной, который задается предикативным центром, и дополняющий его, выраженный с помощью обособления. Таким образом достигается увеличение информативности предложения и текста.

Определение связей и границ различных обособленных конструкций опирается на несколько общих предварительных допущений. Во-первых, связи и границы этих единиц не выходят за пределы предикативной части. Во-вторых, связи и границы могут устанавливаться только слева или только справа от сказуемого.

Логика алгоритма поиска синтаксических связей опорного слова конструкции диктуется языковыми и текстовыми характеристиками этого слова: сочетаемостью и позицией.

Общее правило установления границ обособленных конструкций - указание первого и последнего слова, синтаксически подчиненных /прямо или опосредованно/ опорному слову. Однако для каждой конструкции есть и особые правила. Например, началом поясняющей конструкции считаются специальные слова А ИМЕННО, В ТОМ ЧИСЛЕ и др.

Обособленные причастные обороты приходится разграничивать с препозитивными необособленными причастными оборотами.

Обособленные конструкции включают в себя предложные группы, сочинительные и инфинитивные конструкции, могут усложняться вводными словами, включать в свою очередь обороты и входить в обороты.

Знакам препинания, указывающим на начало и конец обособления, приписывается соответствующая информация.

КРЫЛОВ Ю.К., ПОЛИКАРПОВ А.А.

РЕАЛИЗАЦИЯ НА ПЭВМ СИСТЕМНОГО АНАЛИЗА КОМПЛЕКСА СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ЛЕКСИКИ

I. В докладе делается попытка системного описания комплекса наблюдаемых параметров, характеризующих статистические свойства лексики естественного языка. Постулируется, что для системного описания необходимо введение не менее трех параметров - V, S, U, отвечающих, соответственно, за план выражения (структурно-формальную характеристику) лексического знака, его языковое содержание и его функции, употребление в коммуникации.

На уровне наблюдения, в зависимости от постановки эксперимента и задач исследования каждый из глубинных факторов может быть репрезентирован своим набором наблюдаемых переменных. В синхронном плане свойства лексики определяются заданием закона распределения слов $f(V, S, U)$ в пространстве вышеуказанных факторов. Изменению языка в диахронии соответствует изменение закона распределения в этом пространстве. Предложенный подход позволяет перейти от проводившегося ранее анализа одно-однозначных соответствий к изучению много-многозначных причинно следственных связей между системно-значимыми характеристиками языка.

2. Рассмотрена модель описания статистических свойств лексики, в которой закон распределения $f(V, S, U)$ конкретизирован плотностью, построенной из онтологических соображений на базе семейства гамма-распределений. Экспериментальная проверка модели показала вполне удовлетворительное соответствие теории с экспериментом.

3. В качестве объекта экспериментального исследования использовались две случайные лексические выборки по 10.000 словоупотреблений каждая, полученные из совокупности текстов художественной прозы общей длиной порядка миллиарда словоупотреблений. Сравнение наблюдаемых частот отдельных слов и их объединений в лингвистически содержательные классы показало, что флуктуации этих частот не превосходят чисто случайные.

4. Общий объем объединенного словаря лексем составил 6393 единицы. Каждой лексеме словаря был поставлен в соответствие вектор наблюдаемых переменных, характеризующих ее

а) в плане выражения
   - длиной словарного слова в фонемах;
   - длиной словарного слова в морфемах;
   - длиной словарного слова в слогах;
   - объемом словоизменительной парадигмы слова в числе встретившихся в выборке словоформ;
   - объемом потенциальной (общеязыковой) словоизменительной парадигмы;

б) в плане содержания
   - числом языковых значений по словарю Ожегова;
   - числом языковых значений по МАС;
   - числом языковых значений по ССРЛЯ;
   - количеством синонимов по словарю А.П. Евгеньевой;

в) в функциональном плане
   - частотой данной лексемы в отдельных подвыборках;
   - набором частот встречаемости этой единицы в различных функциональных стилях по данным словаря Засориной;
   - числом функциональных стилей, в которых она встречается в словаре Засориной;
   - числом тематических категорий, которые покрываются данным словом.

Указанная информация реализована в виде базы данных на ПЭВМ IBM - PC/XT и управляется программным комплексом, написанным на языке TURBO-PASCAL. Принятый способ представления данных позволяет расширить как объем исследуемого словаря лексем, так и набор их характеристик.

Проведенные эксперименты позволили проанализировать ряд не рассматривавшихся ранее зависимостей между единицами лексической системы языка, например, изучить характер непосредственной связи между длиной и полисемией лексических единиц.

КРЫЛОВ Ю.К., СЕРГЕЕВ А.Г.

ВОЛНОВАЯ ТЕОРИЯ, ВРЕМЕННЫЕ РЯДЫ И ПРОБЛЕМА СТАТИСТИЧЕСКИХ ОЦЕНОК ЦЕЛОСТНОСТИ ХУДОЖЕСТВЕННОГО ПРОИЗВЕДЕНИЯ

1. Согласно основным положениям волновой теории порождение текста рассматривается как суперпозиция

$$\Psi(x) = \sum_k C_k \Psi_k(x)$$

элементарных случайных процессов $\Psi_k(x)$, развивающихся в пространстве допустимых состояний. Статистические характеристики многомерного процесса "в целом" определяются заданием волновой функции $\Psi(x)$.

Волновые функции $\Psi_k(x)$ соответствующие отдельным состояниям, полагаются ортонормированными на промежутке [a,b]. При этом из условия $\int_a^b \Psi^2(x)dx = 1$ и известного тождества Бесселя $\int_a^b \Psi^2(x)dx = \sum_k C_k^2$ следует равенство $\sum_k C_k^2 = 1$, что позволяет интерпретировать C_k^2 как вероятность возбуждения k-го состояния. Соответственно, если k-е состояние связывается, например, с совокупностью слов, вошедших в текст ровно k раз, то число таких слов в тексте оказывается равно $m_k = C_k^2 \mathcal{L}$, где $\mathcal{L}$ – словарь текста.

2. Для построения конкретной модели порождения текста требуется задать вид функции $\Psi(x)$ и фиксировать систему ортонормированных функций. Рассмотренные ранее модели строились в предположении константного характера $\Psi(x)$, а именно $\Psi(x) = Const = \frac{1}{\sqrt{N}}$

Очевидно, что соответствие теории с экспериментом может быть улучшено, если выбор вида $\Psi(x)$ и $\Psi_k(x)$ осуществить не умозрительно, а на базе предварительного изучения эмпирического материла.

3. Возможности современной вычислительной техники и имеющийся математический аппарат анализа "временных" рядов позволяют провести детальный анализ статики (покрытие) и динамики изменения статистических характеристик индивидуальных текстов на различных уровнях их организации.

Учет специфики рассматриваемых процессов (целочисленные значения компонент временных рядов, спектральное представление по синусам и косинусам нечетных дуг, естественные граничные условия для операторов сглаживания и т.п.) позволил разработать упрощенные вычислительные алгоритмы построения коррелограмм, быстрого преобразования Фурье и сглаживания полученных экспериментальных зависимостей, обеспечивающие достаточную точность вычислений. Особо анализировались возможности представления исходных рядов кумулятивными суммами с последующей аппроксимацией их кусочно-линейными и другими непрерывными функциями.

4. Разработанные вычислительные процедуры были использованы для изучения статистических свойств достаточно представительного массива текстов художественной прозы. Исследовались динамика нарастания объема словаря, плотность покрытия текста словами различных кратностей и словаря в целом. С целью выяснения различий в статистических свойствах отдельных кусков связного текста и целостного произведения, особое внимание уделялось изучению низких частот спектральных представлений процессов.

Установлено, что динамика изменения лингвистически интерпретируемых параметров текстов, действительно, хорошо аппроксимируется тригонометрическими многочленами весьма низких порядков. Полученные предварительные результаты позволяют высказать гипотезу, что в отличие от куска текста произвольной длины целостное произведение "в идеале" должно обладать дополнительной симметрией. Соответственно, если динамика нарастания словаря текста L(X) описывается разложением в ряд $L(x)=\sum_k a_k \cos\frac{(2k-1)\pi x}{2N}+b_k \sin\frac{(2k-1)\pi x}{2N}$ то в качестве меры его целостности может использоваться, например, $\eta=\sum_k b_k^2 / \sum_k (a_k^2+b_k^2)$ либо другие аналогичные отношения.

КУКЕМЕЛК Х.Б., МИКК Я.А.

ПРОГНОЗИРОВАНИЕ ЭФФЕКТИВНОСТИ УЧЕБНОГО ТЕКСТА НА ОСНОВЕ ЕГО МАШИННОГО АНАЛИЗА*

В последние годы все более актуальной становится проблема создания оптимальных по сложности и трудоемкости учебников. Для достижения этой цели необходимы всесторонний анализ и оценка их сложности и трудоемкости. В идеальном случае текст рукописи учебника вводится в ЭВМ и с помощью соответствующих программ определяется контингент учащихся, которому данный учебник доступен.

В основу программ анализа следует принять данные об эффективности разных учебных текстов. Для ее определения Тартуский университет в 1985/86 учебном году провел соответствующий эксперимент в школах республики с русским языком обучения. В эксперименте участвовало свыше 400 учеников. На основе экспериментальных данных мы вычислили процент правильны ответов учащихся по каждому параграфу и их оценку интересности каждого параграфа. Суммированием стантартизированных значений этих показателей мы получили показатель эффективности учебного текста.

Все экспериментальные тексты мы ввели в ЭВМ ЕС-1060 и осуществили морфологический анализ на основе программ, разработанных в Киеве под руководством Н.А. Дарчук. Словам анализируемого текста ЭВМ приписала частотность по данным разных словарей, зафиксированных в ее памяти. В результате машинного анализ текстов были вычислены следующие группы их признаков.

1. Длина текста (количество предложений, слов и т.д.).

2. Процент различных частей речи, использованных в тексте (существительных, прилагательных, числительных, глаголов и т.д.).

3. Различные показатели частотности слов в тексте (частотность по словарю разговорной речи всех слов текста и их грамматических категорий; процент имен существительных, не вошедших в словарь Засориной; повторяемость слов в тексте и т.д.).

---

* На разных этапах работы в ней принимали участие также Г. Алексина, С. Квитко, И. Созин, Т. Боровская, У. Вольмер, О. Оршер, Е. Сивенкова, Л. Томаш и др.

4. Характеристика предложений (процент предложений, длиной до 5 слов, 7 слов; длина именных конструкций и т.д.).

5. Длина используемых в тексте слов.

В результате регрессионного анализа мы получили формулу для оценки эффективности учебного текста, которая выглядит следующим образом:

$$y = 15{,}24 - 5{,}16x_{40} - 0{,}0059x_{43} - 0{,}0498x_{57} + 0{,}482x_{121} + 0{,}0224x_{237} - 2{,}53x_{337} - 0{,}0255x_{344} - 0{,}315x_{371} + 0{,}470x_{419} - 0{,}173x_{430};$$ - где

x_{40} - средняя абстрактность имен существительных в тексте по трехбалльной шкале;

x_{43} - средняя частотность имен существительных текста по словарю учебника физики УШ класса;

x_{57} - процент имен существительных с абстрактностью I;

x_{121} - количество имен существительных в самостоятельной части предложения;

x_{237} - процент имен существительных из словаря 4000 наиболее часто употребляемых в языке слов;

x_{337} - средняя частотность имен прилагательных в тексте;

x_{344} -процент наречий в тексте, не встречающихся в словаре разговорной речи;

x_{371} - процент предлогов в словаре текста;

x_{419} - процент слов в тексте, состоящих из 12 букв;

x_{430} - количество символов на 100 слов текста.

Полученная формула согласуется с известными в теории читабельности положениями: эффективность восприятия текста уменьшается с увеличением в нем абстрактных слов (x_{40} и x_{57}); с увеличением редко употребляемых в языке слов (x_{237} и x_{344}), и символов (x_{430}). Чем больше в тексте слов из учебника физики УШ класса (x_{43}) и чем больше повторяемость в анализируемом тексте имен прилагательных (x_{337}), тем меньше интерес школьников к этому тексту.

Вычисляемый показатель эффективности выше нуля, если анализируемый текстлучше показателей общесоюзных учебников физики для IX и X классов, полученных в 1985 году. Полученный коэффициент множественной корреляции очень высок - 0,97. Это показывает преимущество машинного прогноза эффективности учебных текстов, Следует только текст подготовляемого учебника ввести в ЭВМ. Такой анализ даст ценную информацию для составителя учебника.

КУРЯКОВ В. Н., ПРОНИН Е. Н.

СЕМИОТИЧЕСКИЕ КОДЫ, ТЕКСТ И МАШИННЫЙ ПЕРЕВОД

При переводе научно-технической литературы нередко возникает необходимость переработки не только вербализованной информации, т. е. записанной знаками буквенно-цифрового кода, но и информации, опредмеченной в знаках других семиотических систем. В ряде случаев в рамках одного первичного документа могут использоваться несколько кодов одновременно (например, аэронавигационные карты ИКАО, радиоэлектронные схемы, чертежи и т. д.)

Машинный перевод таких документов (текстов) не может считаться адекватным, если из процесса перевода будет исключена какая-либо из задействованных в нем знаковых систем. Различаем следующие наиболее употребительные семиотические коды: буквенно-цифровой; код математических формул; код химических формул; графические коды; код топографических условных знаков; коды специальных условных знаков и т. д.

Адекватность декодирования-перекодирования текста переводчиком зависит от уровня его компетентности в конкретной сфере/сферах общения и лингвистической компетентности. Для ЭВМ все задействованные в тексте коды равновалентны и, следовательно, информация, содержащаяся в тексте, может быть адекватно репрезентирована в любом удобном для пользователя коде.

Под текстом (с точки зрения машиного перевода) понимаем совокупность знаков семиотических систем, применяемых в данной сфере общения, характеризующуюся целостностью, связностью, смысловой и информационной полнотой и законченностью.

Практическая работа по решению проблемы машинного перевода формализованных поликодовых тестов показывет, что одним из путей является создание универсальных программ, позволяющих переводить заданные специальные коды в буквенно-цифровой и учитывающих различия в информационной емкости различных семиотических кодов.

ЛАВРЕНЕНКО А.В.

ПОЛИПРЕФИКСАЦИЯ И ЕЁ ПРЕДСТАВЛЕНИЕ В МАШИННОМ СЛОВАРЕ

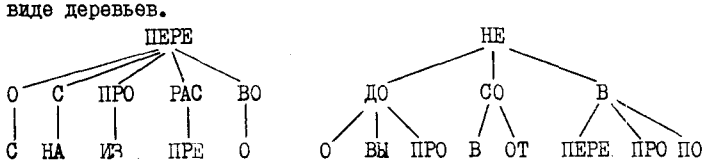
Создание полного машинного словаря предполагает учет как уже фиксированных слов, так и не зафиксированных и тех, которые могут появиться. В НИЛ теоретической и прикладной лингвистики БГУ разрабатывается матричная форма словаря: матрица стандартна по префиксам и финалям и образует словообразовательное гнездо. Такой трёхмерный словарь даёт возможность идентифицировать слова, не заложенные в памяти, и вычислять их семантику. Поиск слов с одним префиксом не представляет трудности, т.к. по предварительным данным префиксов не более 100 с вариантами, а омонимия корней минимальна (имеется в виду случаи неправильного деления: по-днять вместо под-нять).

Определённую трудность при обработке будут представлять слова с несколькими префиксами. Они в основном низкочастотны, но многочисленны и разнообразны. Трудность в том, чтобы наиболее полно представить список двойных, тройных и т.д. префиксов в словаре. Все теоретически возможные сочетания префиксов по 2 можно получить декартовым произведением. Учитывая, что полипрефиксация охватывает в основном русские префиксы, ограничимся анализом таблицы из 841 возможного сочетания (29х29). Для наглядности все варианты префиксов приведены к одному. (При вводе в машину варианты представляются как самостоятельные префиксы, появятся зоны орфографических запретов. При обнаружении опечатки возможно автоматическое её исправление). Из всех теоретически возможных сочетаний реализовано, по данным Обратного словаря русского языка, 247 (в анализе представлены пока наиболее мощные матрицы). Имеется 58 пар взаимнообратных вариантов, или "лево-правых" префиксов. Остальные имеют единственно возможный порядок следования префиксов, т.е. являются "условно левыми" или "условно правыми". Префиксы с одним конкретным значением чаще употребляются в I позиции, например: не - 21:7, без - 17:3, пред - 14:2. Многозначные префиксы одинаково употребительны в I и 2 позициях: по - 18:20, с - 18:21, на - 16:16. Точное распределение слов с двойными префиксами по частям речи пока не сделано, но общая тенденция такова: в основном полипрефиксация представлена в глаголах, в наречиях, образованных от существительных с предлогом, в прилагательных и причастиях, к

которым присоединяется префикс не-, и в образованных от них существительных.

Получение полного списка возможных двойных префиксов с помощью декартова произведения обеспечивает в некоторой степени автоматический анализ новых слов, не введенных в память машины. Например, среди реализованных двойных префиксов нет сочетания "до-пере-". Но в банке данных НИЛ есть слово "доперестроечный", которое должно быть проанализировано и записано в соответствующую матрицу. Только последние накопления новых слов заполнили 7 клеток декартова произведения, из них одна – на оси симметрии: переперестройка. Полная же автоматизация при обработке новых слов возможна только с полными списками корней и финалей.

На уровне двойных префиксов в большинстве случаев нельзя предсказать появление того или иного корня в зависимости от префикса, за исключением некоторых: ни-от-куда, со-с-читать или со-с-тыковать. Такая предсказуемость появляется на уровне тройных префиксов и определяется либо однозначно, либо небольшим перебором на корневой матрице, например: пере-о-смыслить, пред-рас-по-ложенность, пере-рас-пре-делить. Такая предсказуемость означает, что появление новых слов с тройными и более префиксами маловероятно: семантика корня чаще всего ограничивает возможности полипрефиксации; кроме того, среди тройных сочетаний первый префикс обычно имеет конкретную семантику и не допускает расширения слева. Следовательно, нет смысла задавать полный список теоретически возможных тройных сочетаний (с учетом вариантов их более 100 тыс.). Можно ограничиться имеющимся материалом, представив его в виде деревьев.



Для упаковки такой структуры имеется пакет программ /I/.

ЛИТЕРАТУРА

I. Структурный анализ символьных последовательностей. Вычислительные системы, 101. Новосибирск, 1984.

ЛЕСНИКОВ С.В., ЗАГОРОВСКАЯ О.В.

МОДУЛЬНОЕ И СТРУКТУРНОЕ ПРОЕКТИРОВАНИЕ АВТОМАТИЗИРОВАННОЙ ЛЕКСИКОГРАФИЧЕСКОЙ СИСТЕМЫ "Г О В О Р"

В процессе работ над Автоматизированным словарем русских народных говоров (АСРНГ) [2] естественно встал вопрос о едином комплексе лингвистических алгоритмов и программ в виде специализированной автоматизированной лексикографической системы (АЛС). Учитывая многолетний опыт работ по решению задач вычислительной лексикографии в рамках создания Машинного фонда русского языка (МФ РЯ) [1] и располагая в качестве "ядра" АЛС словарно- и тексто-ориентированными подсистемами АЛС "УНИЛЕКС" [3;5], разработчики АСРНГ пришли к выводу: АСРНГ, и в целом Диалектологический подфонд (ДФ) МФ РЯ, нуждаются в собственной диалектологической АЛС. В качестве таковой и разрабатывается АЛС "ГОВОР" [4] в Сыктывкарском государственном университете силами кафедры русского языка и информационного центра.

Для оптимального обеспечения процесса автоматизации диалектологических исследований лингво-программными средствами главными принципами приняты структурность и модульность как программного, так и лингвистического обеспечений при соблюдении принципа независимости лингвистического обеспечения от программного.

Названные принципы определены и самой организацией АЛС "ГОВОР" (меню начального диалога):

1. Описание АЛС "ГОВОР" (HELP).
2. Словарный банк данных (СБД).
3. Текстовой банк данных (ТБД).
4. Связь СБД и ТБД (интерфейс).
5. Копирование (магнитные диск и лента; листинг; экран дисплея).
6. "Пакет" заданий (формирование заданий для выполнения в автоматическом режиме).
7. Библиография.
8. Инструкции. Руководства. Отчеты. (Научно-методическая документация).

Предлагаемое меню режимов работы, иерархически раскрываясь

"сверху-вниз", позволит филологу, обладающему некоторым минимальным уровнем познаний современной вычислительной техники, выбрать требуемый путь компьютерной обработки лексикографических материалов, находящихся на машинных носителях, с последующим постепенным обучением возможностям системы "ГОВОР". АЛС "ГОВОР"предполагает: создание различных словарей (словников) на основе ТБД с определением структуры словарной статьи автоматизированного словаря (АС) в режиме диалога; пополнение СБД материалами ТБД в автоматизированном (интерактивном) режиме; формирование не только дифференциальных, но и полных словарей отдельных говоров с выходом на литературные АС и ТБД.

Вышеназванные принципы создания АЛС "ГОВОР" связаны со структурой самого МФ РЯ. В настоящий момент система "ГОВОР" отрабатывается на материалах одного говора - с.Лойма Прилузского района Коми АССР, результатом чего явится версия АС говора с.Лойма (АСГЛ). АСГЛ входит в состав АС русских говоров территории Коми АССР и сопредельных областей, на котором базируется создание АСРНГ. АСРНГ - основа Словарного подфонда ДФ МФ РЯ. В свою очередь, ДФ МФ РЯ - модуль Академического словарно-грамматического фонда МФ РЯ. При соответствующей доработке АЛС "ГОВОР" может быть использована на всех уровнях МФ РЯ вплоть до ДФ МФ РЯ включительно.

Литература

I. Андрющенко В.М. Концепция и архитектура Машинного фонда русского языка. М.: Наука, 1989. 200 с.

2. Загоровская О.В., Лесников С.В. Автоматизированный словарь русских народных говоров как основа Диалектологического подфонда Машинного фонда русского языка // Третья Всесоюзная конференция по созданию Машинного фонда русского языка. Тезисы докладов. Ч.2. М., ИРЯз АН СССР, 1989. С. 5-7.

3. Колодяжная Л.И. Автоматизированная лексикографическая система УНИЛЕКС. Словарно-ориентированная подсистема. М.: Изд-во МГУ, 1987. 116 с.

4. Лесников С.В. Архитектоника Автоматизированной лексикографической системы "ГОВОР" // Третья Всесоюзная конференция по созданию Машинного фонда русского языка. Тезисы докладов. Ч.2. М., ИРЯз АН СССР, 1989. С. 7-8.

5. Мошкович Ж.Г. Автоматизированная лексикографическая система УНИЛЕКС-2. М.: Изд-во МГУ, 1989. 107 с.

МАЛОВ А.В.

СКАНИРУЮЩАЯ ПРОГРАММА ДЛЯ ОПРЕДЕЛЕНИЯ ПАРАМЕТРОВ РАНГОВЫХ ПОЛИСЕМИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

Разработана программа для определения параметров ранговых полисемических распределений лексики по аппроксимирующим формулам

$$(p_i + A)(i^{\gamma} + A) = C$$

$$lg^{\alpha} p_i + lg^{\alpha} i^{\gamma} = K,$$

где p_i - степень полисемичности слова в рангом i, A, α, γ, C, K - константы.

Формулы получены на основе некоторых феноменологических соображений.

Сканирование осуществляется по трем параметрам, ответственным за максимальное значение кривой по оси полисемии, за наклон и за выпуклость графика. На материале словарей русского и английского языков было установлено, что программа позволяет достичь высокого качества аппроксимации по предложенным формулам, что обеспечивает получение объективной информации об особенностях распределений конкретных толковых словарей, в частности, вызванных типологической принадлежностью данного языка (синтетизм/аналитизм) и типом словаря (краткий, средний, большой). При использовании предположения о большей статистической достоверности средней части распределения по сравнению с его концевыми участками возможно осуществление аппроксимации распределения с большим учетом именно средней части распределения.

В качестве минимизируемого показателя близости эмпирического и теоретического распределения использовалась площадь фигуры между двумя соответствующими графиками в билогарифмическом масштабе.

Программа реализована на ЭВМ ЕС-1061. В настоящее время производится отладка аналогичной программы на ПЭВМ IBM PC/XT.

МАЛЬКОВСКИЙ М.Г., БОГАЧЕВ Д.Н
ВИТКАЛОВА Е.В.

ОБРАБОТКА АББРЕВИАТУР И СЛОВ СО СЛОЖНОСОСТАВНОЙ СВОБОДНООБРАЗУЕМОЙ СТРУКТУРОЙ В СИСТЕМАХ АВТОМАТИЧЕСКОГО АНАЛИЗА ЕЯ-ТЕКСТА

Излюбленным приемом авторов текстов научной тематики стало введение сокращений вместо употребляемых терминологических словосочетаний и использование этих новых, неслсварных объектов в дальнейшем повествовании. Включение свобсднообразуемых сложных слов нецелесообразно ввиду неограниченности их числа.

В докладе мы расскажем, как указанные объекты обрабатываются в системе ЛИНАР, литературном научном редакторе, создаваемом на кафедре АЯ ВМК МГУ под руководством М.Г. Мальковского.

ЛИНАР включает два словаря сокращений: общий и предметной области. В процессе анализа текста формируется список сокращений текста вместе с расшифровками и список неопознанных аббревиатур. При встрече с аббревиатурой в тексте, система по ряду признаков определяет, присутствует ли в предложении ее дефиниция. Если расшифровка должна быть, то она выделяется. Далее аббревиатура ишется в списках сокращегий текста, словарях сокращений предметной области и общем. На выходе имеем список сокращений текста вместе в расшифровкой.

Для обработки слоговых аббревиатур (мосавтодорстрой, облремонтстройтрест и т.п.) и слов сложносоставной, свободнообразуемой структуры (восьмиэтажный, пионервожатый) в систему вводится словарь наиболее продуктивных при словообразовании абброморфем (мос-, ком-, одно и т.п.). В словарную статью словаря абброморфем включается следующая информация: полная форма, признак опорности, грамматические и семантические характеристики, передаваемые абброморфемой сложному слову в случае опорности. Разработаны алгоритмы разбиения сложного слова на части и синтеза морфологических и семантических характеристик сложного слова по частям его составляющим.

МАНАСЯН Н.С.

ОБ ИСПОЛЬЗОВАНИИ ДОКУМЕНТАЦИИ НА МАШИННЫХ НОСИТЕЛЯХ В ЛИНГВОСТАТИСТИЧЕСКИХ ЦЕЛЯХ

Предельным случаем описания текста лингвостатистическим путем является его описание "вглубь". Материалом для такого описания может послужить единый и достаточно большой корпус текстов на одном языке и принадлежащих одной тематике. Последняя должна быть достаточно узкой. Условия достаточно большой выборки и по возможности узкой, неделимой тематики обеспечивают высокую надежность полученных результатов. Таким материалом послужили тексты на русском языке общей длиной в 750 тыс. словоупотреблений и представляющие собой документацию ЭВМ. В целях исследования количественных закономерностей было проведено поуровневое рассмотрение встречаемости лингвистических единиц в данном тексте. Перечень нижеприводимых признаков отражает последовательность поуровнего анализа лингвистических единиц рассматриваемого научно-технического текста.

Символы

I.I. Встречаемость букв.

I.2. Спектровое распределение букв.

I.3. Начальные и конечные буквы слова.

Слова

2.I. Встречаемость слов.

2.2. Спектровые распределения лексем.

2.3. Спектровые распределения словоформ.

2.4. Частные распределения для лексем.

2.5. Распределения слов по длине в буквах.

2.6. Устойчивые сочетания слов.

Предложение

3.I. Начальные и конечные слова в предложении.

3.2. Распределение предложений по длине в словах.

Абзац

4.I. Распределение абзацев по длине в предложениях.

4.2. Начальные слова в абзацах.

Текст разбивался на отрезки длиной в 500 слов. Слово - последовательность русских букв, заключенная между двумя символами, не являющимися русскими буквами.

МОИСЕЕВА С.А.

УСТАРЕНИЕ МНОГОЗНАЧНОСТИ ЛЕКСИЧЕСКИХ ЕДИНИЦ В РАМКАХ ТИПОВЫХ КОНТЕКСТОВ

При автоматической переработке текста и машинном переводе одним из еффектных способов снятия многозначности слова является актуализация его значений по типовым контекстам.

В слове как основной единице языка сосредоточены лексические и грамматические значения. В зависимости от типов лингвистических знаков, взаимодействующих внутри типовых контекстов, последние делятся на две группы: I/ грамматические, в которых выявление значения многозначных слов осуществляется на семантико-синтаксическом уровне, и 2/ лексико-семантические типовые контексты. Актуализация значения многозначных слов в них происходит через взаимодействие семантики входящих в данный типовой контекст лексических единиц.

Механизм актуализации значения многозначного слова в грамматических и лексико-семантических типовых контекстах различен. Грамматические типовые контексты французских многозначных существительных выявляются на синтаксическом и морфолого-синтаксическом уровнях. Типовой контекст, в котором синтаксическое /линейное описание/ исчерпывающе дифференцирует значение существительного, является синтаксическим типовым контекстом. Когда для актуализации многозначного слова наряду с синтаксическими подключаются и морфологические характеристики имени существительного, то они образуют морфолого-синтаксический типовой контекст.

Устранение многозначности на лексико - семантическом уровне - более сложный и тонкий способ актуализации значения слова. Для выявления значения неоднозначного существительного выделяются семантические классы как самого существительного, так и единиц его окружения /глаголов, существительных и прилагательных/, строятся матрицы сочетаемости лексических единиц в типовом контексте, на основе которых осуществляется снятие неоднозначности слова. В результате анализа всей полученной информации на лексико - семантическом уровне выделяются следующие виды типовых контекстов: именные предложные, именные беспредложные, адъективные и глагольные типовые контексты. В качестве примера в докладе приводится анализ существительного 'terme' и выявление его значений во всех вышеперечисленных типовых контекстах.

МОРОЗОВА Е.Н.

ЯЗЫКОВОНЕЗАВИСИМЫЙ СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР

Синтаксический анализатор(СА) предназначается для автоматического синтаксического анализа предложений на естественном языке и создается на базе модели синтаксиса, разрабатываемой Ю.С.Мартемьяновым. СА является системой поверхностно-синтаксического анализа и имеет на входе предложения естественного языка, т.е. цепочки словоформ, а на выходе - синтактико-лексические структуры (деревья зависимостей), поставленные в соответствие цепочкам словоформ предложений. Началу и концу каждой стрелки дерева приписываются согласованные(парные) синтаксические категории, определяющие тип связи между словоформами. Выходные синтактико-лексические структуры строятся на основе словаря, включающего в себя морфологические и синтаксические сведения, и ряда таблиц, содержащих правила анализа. СА является языковонезависимым в том смысле, что для перехода от анализа предложений на одном естественном языке к анализу предложений на другом естественном языке достаточно сменить словарь и, возможно, таблицы правил, в то время как алгоритмическое и программное обеспечение остается неизменным. Проведены эксперименты по анализу русских и французских предложений.

Словарь анализа на данном этапе разработки СА является словарем словоформ. Синтаксическая информация в словарных статьях представлена в виде упорядоченного набора синтаксических категорий, определяющих роль словоформы в структуре, так называемых поисковых признаков, задающих активность слова в поиске пары, направление поиска слова, связанного с данным, отношения главенствования/зависимости между словами в паре, тесноту связи - характеристику вероятного расстояния до партнера, а также сведений об управлении/согласовании по синтаксическим категориям и поисковым признакам.

Алгоритмическое обеспечение СА состоит из 5 основных блоков. Первый из них - общий алгоритм - реализует анализ "слева направо", т.е. строит рабочее пространство (РП), начиная с первых двух слов предложения и добавляя очередное слово после того, как для данного РП выполнены все возможные действия по установлению связей между содержащимися в нем словами; и так до тех пор, пока предложение не будет исчерпано. Общий алгоритм организует также обращения к остальным алгоритмическим

блокам. Эти блоки, а именно: блок установления связей, блок обработки конфликтных ситуаций, блок категоризации и блок изменения информации, - реализованы как интерпретаторы языков, на которых записаны правила в соответствующих таблицах. Вместе с правилами заданы условия их применения, проверка удовлетворения этих условий проводится либо в общем алгоритме, либо в вызванном им блоке.

В блоке установления связей происходит предварительное установление связи между двумя словами. В блоке обработки конфликтных ситуаций осуществляется проверка, не ведет ли добавление новой связи в строящуюся синтактико-лексическую структуру к нарушению ее древесности; если это так, необходимо отменить либо вновь устанавливаемую, либо одну из ранее существовавших связей. В этом же блоке производится обработка непроективных структур. Блок категоризации служит для приписывания синтаксических категорий началу и концу вновь установленной связи, причем приписываемые категории либо извлекаются из словарных статей, либо вычисляются по правилам, содержащимся в специальной таблице; здесь же могут проводиться дополнительные проверки правильности новой связи. В блоке изменения информации по результатам установления новой связи меняются значения некоторых переменных, связанных со словарными статьями и строящейся синтактико-лексической структурой, что позволяет управлять процессом анализа с целью его ускорения за счет исключения лишних альтернатив.

Л И Т Е Р А Т У Р А

I. Мартемьянов Ю.С., Морозова Е.Н. Свойства синтезирующего описания в применении к анализу текста. // Машинный фонд русского языка. Предпроектные исследования. М.,ИРЯ,1988.

МОШКОВИЧ Ю.Г.
АНОШКИН Е.А.

ОДИН СПОСОБ ЛЕММАТИЗАЦИИ ДЛЯ АВТОМАТИЧЕСКОГО КОНКОРДАНСА

Под лемматизацией мы подразумеваем приведение текстовых форм слов (словоформ) к словарным (леммам): существительного - к именительному падежу единственного числа; прилагательного - к форме именительного падежа единственного числа мужского рода; всех форм глагола, включая причастия и деепричастия, - к инфинитиву и т.п.

В Институте русского языка АН СССР реализована лемматизация на ПЭВМ типа IBM с использованием словаря словоформ. В результате лемматизации каждой словоформе исходного текста приписывается словарная форма (лемма) и грамматическая помета (часть речи). Словоформы, не найденные в словаре, анализируются по аналогии с имеющимися в словаре.

После проверки и коррекции результатов лемматизации словарь может быть пополнен материалом обработанного текста. Таким образом происходит как бы "обучение" программы лемматизации и в то же время настройка ее на определенную тематическую область.

Лемматизации может быть использована в различных системах обработки текстов. Ниже приводится пример подключения ее к пакету ETC, который предназначен для получения автоматических конкордансов на персональных компьютерах. Элементарной единицей запроса на выдачу контекста является словоформа.

В пакете предусмотрены некоторые дополнительные файлы, один из которых - так называемый "тезаурус". Он представляет собой множество каких-либо групп словоформ, снабженных заголовком, и создается вне системы ETC.

Именно этот механизм - "тезаурус" - был реализован нами для использования результатов лемматизации в пакете ETC. После осуществления лемматизации и коррекции результатов формируется "тезаурус", соответствующий требованиям пакета ETC. Каждую группу "тезауруса" составляют словоформы, относящиеся к одной лексеме, комментарием у них является грамматическая помета и частота, а заглавным словом группы - лемма. Такой "тезаурус" позволяет собрать вместе все встретившиеся в тексте словоформы одной лексемы, иметь перед глазами такой список и переходить от леммы к словоформам, а от словоформы - к контекстам.

НЕЛЮБИН Л. Л

ИНЖЕНЕРНО-ЛИНГВИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ОТРАСЛЕВЫХ ПОДЪЯЗЫКОВ

1. Комплексный вероятностно-статистический и информационный анализ подъязыка с относительно конечным числом состояний позволяет создать базовый подъязык и осуществить его реализацию в инженерно-лингвистической модели в интересах вероятностного исполтьзования полученных результатов в АСУ.

2. Методика построения заключается в проведениии поэтапного моделирования. Созданная на основе функционально-коммуникативного анализа программа выборки, представляющая собой функционально-коммуникативную модель организационнойй структуры подъязыка, позволяет провести эффективное вероятностно-статистическое обследование корпуса, результирующееся в создании серии частотных словарей (ранговых, алфавитных, обратных), которые являются моделью подъязыка, а проведение информационных измерений совокупности позволяет составить ее информационную модель. На основании информационно-статистических данных осуществляется отбор лексики в базовый подъязык, представляющий ту лингвистическую модель, которая получает машинную интерпретацию и преобразовывается в машинный автономный отраслевой словарь (МАОС), составляющий в совокупности с соответствующим инженерным обеспечением инженерно-лингвистическую модель подъязыка (ИЛМ).

3. Реализация частной методики алгоритмической обработки информации выражается в построении компактной и автономной (независимой) инженерно-лингвистической модели, которая может быть использована уже в настоящее время для решения определенных задач по обработке текстов определенной тематики в условиях одноязычной и двухязычной ситуации, как для систем машинного перевода (МП), так и систем перевода с помощью машины (ПМ).

4. Анализ известных нам (действующих и разрабатываемых) систем автоматизированной переработки информации показывает, что существующие перспективные инженерно-лингвистические модели МП и ПМ могут быть успешно реализованы, если они будут решаться как комплексная инженерно-лингвистическая задача.

5. Прагматический аспект инженерно-лингвистического моделирования осуществлен в иконическом режиме на ПЭВМ и ЭВМ для систем МП и ПМ.

ОСТАПЕНКО В.Е.

АЛГЕБРАИЧЕСКОЕ МОДЕЛИРОВАНИЕ ТЕКСТА

Попытки представить язык в виде исчисления неоднократно предпринимались в 60-е годы нашего столетия. При этом под языком понималось конечное или бесконечное множество предложений, каждое из которых имеет конечную длину и построено с помощью операции соединения из конечного множества элементов. Это определение включает как естественные, так и искусственные языки логики и программирования.

I. Процедура построения такого исчисления начиналась с анализа некоторой конечной совокупности речевых цепочек с целью получения их структурного описания, позволяющего в дальнейшем синтезировать практически ничем не ограниченное число грамматически правильных предложений. Однако предпринятые попытки алгебраизации языка методами неколичественной математики показали, что получаемые исчисления порождают большое количество правильных и одновременно бессмысленных предложений, которые необходимо отфильтровать с помощью ограничений.

2. Попытка алгебраизации языка с помощью количественной математики может быть реализована путем построения алгебраической модели текста (АМТ), которая в математических терминах определяется как комбинаторная конфигурация (КК) или система алгебраических уравнений (САУ). Казалось бы перевод задач на язык алгебры, их решение на этом языке, а затем обратный перевод только усложняют дело. В действительности такой путь оказывается весьма выгодным, а порой и единственно возможным при моделировании языка в компьютере.

3. Процедура построения КК или АМТ начинается с описания распределения лексических единиц (ЛЕ) в тексте с помощью семейства степенных функций, число которых в семействе для каждой ЛЕ зависит от частоты ее употребления в тексте. На этом этапе моделирования текста КК представляет собой неупорядоченное и неструктурированное множество ЛЕ, которое малоинформативно относительно их значимости как в системе языка, так и в тексте. Для того, чтобы повысить лингвистическую значимость полученной КК, необходимо сначала упорядочить полученный инвентарь ЛЕ по убыванию частоты их употребления в тексте, затем

разбить это упорядоченное множество на подмножества, включающие равночастотные языковые элементы или ЛЕ с приблизительно равной частотой таким образом, чтобы мощность получаемых подмножеств была практически одинаковой. Подобное структурирование КК позволяет отделить в ней зону, содержащую в основном высокочастотные термины макрополя, отражающие общую тематику текста, от средне- и низкочастотной зоны, содержащей такие термины, которые определяют тематику его фрагментов. На очередном этапе структурирования КК проводится упорядочение ЛЕ каждого подмножества по убыванию значений показателя степени соответствующей функции их распределения в тексте, в результате чего каждое подмножество приобретает лингвистически значимую четкую или приемлемо четкую структуру, где на периферии группируется терминологическая лексика, в то время как ядро этих подмножеств в основном состоит из общеупотребительных ЛЕ. Такая структура КК позволяет проводить формальную классификацию ЛЕ с целью дальнейшего выделения наиболее информативных фрагментов текста, которые, очевидно, характеризуются сгустками терминологии или какими-либо другими ЛЕ и их сочетаниями, релевантными относительно того, или иного круга пользователей.

4. Процедура извлечения необходимой информации из текста может быть реализована как с помощью слабо направленного перебора терминологических ЛЕ, так и решением систем алгебраических уравнений, где стандартные правила решения задают способ построения конструкции из элементов КК или АМТ. В этом случае КК автоматически превращается в исчисление, где ее элементы - ЛЕ выполняют роль зависимых переменных, параметры степенной функции их распределения в тексте - роль ограничений, моделирующих контекстную зависимость грамматики, а совокупность КК и АМТ с правилами решения систем уравнений алгебраически моделирует процесс порождения текста в целом.

ПАВЛОВИЧ Н.В., ЧЛЕНОВА С.Ф.

СЛОВАРЬ РУССКИХ ПОЭТИЧЕСКИХ ОБРАЗОВ НА ЭВМ

Словарь русских поэтических образов создается по текстам русской художественной литературы 18 - 20 вв. В настоящее время с помощью системы "Диабас", разработанной в Институте проблем управления АН СССР, на ЭВМ СМ-4 создана экспериментальная база словаря, включающая около 6000 словарных статей.

Под поэтическим образом понимается отрезок текста (от одного слова до нескольких строф или предложений), в котором сближаются (отождествляются) несходные (семантически далекие, несовместимые, в том числе противоположные и противоречащие понятия). Примеры образов: холодожара (сближаются холод и жара), звезд булавки золотые (сближаются звезды и булавки), пуховый платок снегопада (снег и пух, снег и ткань) и т.д.

Главное содержание словарной статьи поэтического образа составляет информация и парадигмах, или моделях, реализованных в этом образе. Парадигма, или модель образа есть устойчивый семантический инвариант вида X $\rightarrow$ У, где X и У - инварианты отождествляемых имен, предикатов, ситуаций, а стрелка указывает направление отождествления. На поверхностном языковом уровне парадигма реализуется в ряде сходных образов. Примеры парадигм (моделей) образов: свет $\rightarrow$ ткань (риза лучей, белой зари рукав, заря златотканная, одежды вечера, саваны утра, рассвет кафтаны опустил), глаза $\rightarrow$ драгоценное (агатовоокая дева, ока холодного жемчуг, дорогие каменья его глаз переливались и таяли, глаза-рубины).

Словарная статья образа содержит 32 поля (среди них - составные и такие, которые представляют собой повторяющиеся группы). Система "Диабас" позволяет вести исследовательскую работу со словарем, т.е. получать ответы на запросы о наличии и количестве образов с любыми комбинациями признаков, указанных в полях словарной статьи. Например, можно получить все образы с такой-то моделью или все образы с заданным списком моделей, либо все образы, где левый член модели, например, "растение $\rightarrow$ сосуд" выражен словом "тополь", а правый - словом "чаша". Система выдает также информацию типа "гистограммы", т.е. список всех разных значений признака с частотами. Так, например, можно получить список всех лексических моделей образов словаря с их частотами.

Ниже приводятся образцы словарных статей словаря образов.

ПОРЯДКОВЫЙ НОМЕР : 00015
ТЕКСТ ОБРАЗА : ОН БЫЛ КАК РЫЖАЯ ГРОЗА.
ГОД НАПИСАНИЯ : 1976
АВТОР : ЛЕВАНСКИЙ
НАЗВАНИЕ ПРОИЗВЕДЕНИЯ : СКАЗКА ПРО КАРЛИКА И ВЕЛИКАНА
ИЗДАНИЕ : ШАРОДЕЙСТВО
СТРАНИЦА : 16
КЛЮЧЕВЫЕ СЛОВА : ГРОЗА ДРАКОН

*** СИСТЕМА "Д И А Б А С" *** *ИАТ, МИНПРИБОР, АН СССР

14:36 5 ОКТ. 1987 Г.

ОБЩЕЕ ЧИСЛО МОДЕЛЕЙ : 1
ЧИСЛО ЛЕКСИЧЕСКИХ МОДЕЛЕЙ : 1
ОБЩИЙ ВИД ЛЕКСИЧЕСКОЙ МОДЕЛИ : СУЩЕСТВО ---> СТИХИЯ
ОБРАТИМОСТЬ : ОБ
ЛЕВЫЙ ЧЛЕН МОДЕЛИ : ДРАКОН
ПРАВЫЙ ЧЛЕН МОДЕЛИ : ГРОЗА

ПОРЯДКОВЫЙ НОМЕР : 00016
ТЕКСТ ОБРАЗА : <...>ОГНЕВОЙ КОЖУРОЙ АБАЖУРА
ГОД НАПИСАНИЯ : 1956
АВТОР : ПАСТЕРНАК
НАЗВАНИЕ ПРОИЗВЕДЕНИЯ : БЕЗ НАЗВАНИЯ
ИЗДАНИЕ : МС
СТРАНИЦА : 358
КЛЮЧЕВЫЕ СЛОВА : КОЖУРА АБАЖУР ОГНЕВОЙ

СЛОЖНОСТЬ ОБРАЗА : СЛ
ОБЩЕЕ ЧИСЛО МОДЕЛЕЙ : 3
ЧИСЛО ЛЕКСИЧЕСКИХ МОДЕЛЕЙ : 3
ОБЩИЙ ВИД ЛЕКСИЧЕСКОЙ МОДЕЛИ : АБАЖУР ---> ПЛОД
ЛЕВЫЙ ЧЛЕН МОДЕЛИ : АБАЖУР
ОБЩИЙ ВИД ЛЕКСИЧЕСКОЙ МОДЕЛИ : СВЕТ ---> ОГОНЬ
ОБРАТИМОСТЬ : ОБ
ПРАВЫЙ ЧЛЕН МОДЕЛИ : ОГНЕВОЙ
ОБЩИЙ ВИД ЛЕКСИЧЕСКОЙ МОДЕЛИ : ПЛОД ---> ОГОНЬ
ЛЕВЫЙ ЧЛЕН МОДЕЛИ : КОЖУРА
ПРАВЫЙ ЧЛЕН МОДЕЛИ : ОГНЕВОЙ

ПОРЯДКОВЫЙ НОМЕР : 00017
ТЕКСТ ОБРАЗА : <...>В ОКНЕ СКУЧАЕТ РОЗА АБАЖУРА
ГОД НАПИСАНИЯ : 1983
АВТОР : АХМАДУЛИНА
НАЗВАНИЕ ПРОИЗВЕДЕНИЯ : ЛУНА ДО УТРА
ИЗДАНИЕ : ТАЙНА
СТРАНИЦА : 27
КЛЮЧЕВЫЕ СЛОВА : РОЗА АБАЖУР
ОБЩЕЕ ЧИСЛО МОДЕЛЕЙ : 1
ЧИСЛО ЛЕКСИЧЕСКИХ МОДЕЛЕЙ : 1

ОБЩИЙ ВИД ЛЕКСИЧЕСКОЙ МОДЕЛИ : АБАЖУР ---> ЦВЕТЕНИЕ
ЛЕВЫЙ ЧЛЕН МОДЕЛИ : АБАЖУР
ПРАВЫЙ ЧЛЕН МОДЕЛИ : РОЗА

ПЕКАРСКАЯ Л.А.

АППАРАТУРНОЕ ОБЕСПЕЧЕНИЕ И АРХИТЕКТУРА СИСТЕМЫ ОБУЧЕНИЯ ГРАММАТИЧЕСКОЙ ПРАВИЛЬНОСТИ РЕЧЕВЫХ ВАРИАНТОВ НА БАЗЕ ПЕРСОНАЛЬНЫХ ЭВМ ЯМАХА

Принципиальной для задач обучения является возможность персональных ЭВМ /ПЭВМ/ объективировать фрагменты мышления в виде языковых и речевых единиц на экране и легко манипулировать этими единицами.

Современные автоматизированные обучающие системы /АОС/, реализованные на ЭВМ и работающие в режиме персональных компьютеров, принципиально меняют методику освоения языковой системы, в том числе и методику обучения грамматической правильности.

В докладе содержится описание созданной в Горьковском государственном педагогическом институте иностранных языков им.Н.А.Добролюбова АОС правильности употребления глагольных форм /для английского языка/ и синтаксических вариантов /для русского языка/.

АОС работает в режиме "меню", что позволяет предъявлять обучаемому различные варианты, из которых осуществляется выбор, что, в свою очередь, инициирует либо повторное предъявление задачи, либо предъявление нового "меню", включающего проверочные варианты. Метод "меню" является эффективным средством реализации на ЭВМ коммуникативных связей "человек-машина", позволяющим формировать и закреплять актуальные и грамматически правильные языковые сущности и связи.

Программный комплекс на языке БЭЙСИК реализован на ПЭВМ ЯМАХА. и состоит из ряда модулей, осуществляющих последовательные операции анализа лексических единиц. Модульная архитектура программного комплекса обеспечивает взаимозаменяемость, доступность и простоту использования программ.

АОС используется для обучения школьников, студентов, стажеров и аспирантов - в курсах "Современный русский язык", "Английский язык", "Основы информатики". Приводятся листинги программ и распечатки дисплейных сообщений.

Методика обучения с применением ПЭВМ позволяет повысить эффективность усвоения грамматических норм, вести обучение в режиме оптимального "человеко-машинного" диалога, обеспечивающего необходимую психологическую комфортность.

ПЕРЕБЕЙНОС В.И.

СТИЛЕРАЗЛИЧИТЕЛЬНАЯ СПОСОБНОСТЬ ЗОН СВЯЗЕЙ

Как было показано в предыдущих работах, зона связей слова /ЗСС/ является основой выделения минимальных текстовых единиц / 2 /, установления границ термина / I /, а также диагностики синтаксической структуры текстового предложения / 3 /. Помимо этого зоны связей обладают стилеразличительными способностями.

В качестве параметров стилей служат черты самих ЗСС /распределение длины, место ЗСС в предложении и др./, а также черты текстовых единиц, порождаемых взаимодействием ЗСС: количество различных уровней связности в текстах разной стилевой принадлежности/, а также их распределение по позициям в предложении, распределение частоты уровней связности, расстояние между пиками связностей в предложении и пр.

Выделяются индивидуализирующие характеристики, отграничивающие некоторый стиль от всех других сопоставляемых стилей, а также параметры, позволяющие группировать тексты по общности их характеристик.

ЛИТЕРАТУРА

I. Дарчук Н.П. Определение границ термина в тексте. // Материалы Всесоюзной конференции "Текст.Термин.Словарь", состоявшейся 12-14 сентября 1989 г. в г.Киеве. - Киев: Наук.думка.

2. Перебейнос В.И. Типология минимальных единиц текста //Linguistica , Вып.838, Тарту, 1988, с.83-93.

3. Перебейнос В.И. Зоны связей как средство определения структуры предложения в тексте //Материалы Всесоюзной конференции, состоявшейся 12-14 сентября 1989 г. в г.Киеве. - Киев: Наук.думка.

ПИЕЛЬ Е.

ТЕКСТОВАЯ БАЗА ДАННЫХ КАК ЭМПИРИЧЕСКАЯ БАЗА КОМПЬЮТЕРНОЙ ЛЕКСИКОГРАФИИ

I. Лексикографическая база данных может быть построена как производная от текстовой базы данных.

2. Преимущества такого подхода состоят в автоматизированном выведении словника и получении контекстов употребления.

3. Работа по схеме текст - словник - текст ведется в отделе математической лингвистики Института языка и литературы им. А.Упита АН ЛатвССР на материале различных функциональных стилей, что позволяет получать разнообразные данные о функционировании единиц различных уровней, а также их квантитативные характеристики. Такой подход дал возможность получить эмпирические распределения языковых единиц [см.I, например,] и дать им теоретическую оценку [4], а также позволил строить вероятностные модели организации лексики текста [2;3].

4. В перспективе предусматривается расширение лексикографической базы для создания машинной версии словаря латышского языка.

ЛИТЕРАТУРА

I. Пиель Е.Ш. Статистика графем латышских художественных и научно-технических текстов // Известия АН ЛатвССР.-1988.-№II.-С.73-83.

2. Скляревич А.Н., Якубайтис Т.А. Возможное обоснование частотных закономерностей лексики текстов. - Рига: ИЭВТ, 1984. - 62 с.

3. Скляревич А.Н., Якубайтис Т.А. Отыскание определяющего параметра частотных закономерностей лексики текстов. - Рига: ИЭВТ, 1985.-65 с.

4. Якубайтис Т.А. Части речи и типы текстов. - Рига: Зинатне, 1981.- 248 с.

ПИОТРОВСКИЙ Р.Г., ПИОТРОВСКАЯ К.Р.,
ГРИГОРЬЕВА Т.В., ТИХОНОВА О.О.,
ЮСУПОВА Ш.Х.

АВТОМАТИЧЕСКАЯ СИСТЕМА ПОМОЩИ ПРИ РАБОТЕ С ИНОЯЗЫЧНЫМ ТЕКСТОМ

Строится многофункциональная система TUTSY для работы с учебным, научно-техническим и деловым текстом на базе системы машинного перевода SILOD=MULTIS. Дидактическая адаптация проводится для франко-русской и немецко-русской версий системы.

Основная идея состоит в том, чтобы, совместив возможности автоматического словаря, грамматического пособия, средств обнаружения орфографических ошибок в тексте, подготовки и редактирования текста, предоставить непрофессиональному переводчику-инженеру, научному работнику-нелингвисту, студенту и школьнику - эффективное средство для работы с иноязычным учебным текстом.

На первом этапе создается лексико-грамматическое компьютерное пособие состоящее из:

- морфолого-синтаксической и семантической справки, которая включает конкретную грамматическую информацию о возможностях актуализации конкретной лексической единицы и устранения омонимии,
- таблиц, дающих обобщенную информацию по запрашиваемой грамматической категории,
- справочника по терминологии, с краткими сведениями по грамматике данного подъязыка и правилами работы с системой.

База данных для морфолого-синтаксической справки основывается на информации словарной статьи автоматического словаря SILOD=MULTIS и организована в виде набора файлов прямого доступа.

Система реализуется на персональных компьютерах класса IBM/PC/AT.

ПОЛИКАРПОВ А.А., САПРЫКИН П.М., СИЛЬНИЦКАЯ Г.В.

КОЛЛИНГ: СИСТЕМА КОЛИЧЕСТВЕННОГО АНАЛИЗА СВЯЗЕЙ МЕЖДУ ХАРАКТЕРИСТИКАМИ ЛЕКСИЧЕСКИХ ЕДИНИЦ

Коллективом лингвистов Смоленского педагогического института под руководством Г.Г.Сильницкого был сформирован список из 410 английских глаголов, полученных путем 5%-ной случайной выборки из словаря "Concise Oxford English Dictionary" и проведена работа по приписыванию им 70 различных фонологических, морфологических, синтаксических и семантических признаков с последующим исследованием корреляционных связей между ними [Сильницкий и др., 1987].

На основе дополнительных исследований нами был расширен список приписываемых глаголам признаков. Дополнительные признаки относятся к словообразовательным, хронологическим, стилистическим, полисемическим, омонимическим характеристикам слов.

Для исследования силы связи между этими признаками и между словами Сапрыкиным П.М. была разработана диалоговая система программ КОЛЛИНГ, реализованная на языке Turbo Pascal для ПЭВМ IBM PC/XT. В качестве критерия оценки связанности использовалось *дистрибутивное расстояние* между признаками (словами), вычисляемое по следующей формуле:

$$\mathfrak{D}(A,B) = 1 - \frac{\sum_{i=1}^{N} (A_i \cap B_i)}{\sum_{i=1}^{N} (A_i \cup B_i)}$$,где

A_i – индикаторная функция наличия слова i с положительно определенным признаком A (или, в случае вычисления дистрибутивного расстояния между словами, признака i в слове A),

B_i – индикаторная функция наличия слова i с положительно определенным признаком B (или, в случае вычисления дистрибутивного расстояния между словами, признака i в слове B),

N – число слов (во втором случае – число признаков),

Механизм вычисления *дистрибутивного расстояния* между признаками (словами) принципиально состоит в выявлении отношения числа общих слов (признаков) для данной пары признаков (слов) к числу всех слов (признаков), на которые распространяются данные признаки (слова). *Дистрибутивное расстояние* варьируется в пределах [0 .. 1].

Выбор *дистрибутивного расстояния* в качестве критерия оценки связан с тем, что для данного слова наличие любого признака может быть определено либо положительно, либо отрицательно.

Для повышения точности анализа и устранения "*шумов*", вызываемых случайной флуктуацией наличия признаков у слов, было введено понятие *веса слова* (признака) или коэффициента его уникальности, вычисляемого по формуле:

$$\mathfrak{V}j = \frac{1}{\sum_{t=1}^{M} P(At,j)}$$,где

$\mathfrak{V}j$ - *вес* j-го слова (признака),
M - число признаков (слов),
$P(At,j)$ - индикаторная функция наличия At-го признака в j-ом слове

С учетом введения *весов слов* (признаков) *дистрибутивное расстояние* между признаками (словами) вычисляется следующим образом:

$$\mathfrak{D}(A,B) = 1 - \frac{\sum_{t=1}^{N} \mathfrak{V}_t (A_t \cap B_t)}{\sum_{t=1}^{N} \mathfrak{V}_t (A_t \cup B_t)}$$

С целью повышения степени достоверности результатов дистрибутивно-статистического анализа каждому *дистрибутивному расстоянию* признаков (слов) ставилась в соответствие величина, названная авторами *коэффициентом достоверности*, вычисление которой производилось следующим образом:

$$\mathfrak{D}(A,B) = \frac{\sum_{t=1}^{N} (A_t \cup B_t)}{N}$$

Коэффициент достоверности следует учитывать, когда для двух пар признаков(слов) *дистрибутивное расстояние* оказалось одинаковым. В этом случае *коэффициент достоверности* будет больше у той пары признаков (слов), у которой имеется большее количество слов (признаков), обладающих этой парой признаков (слов) .

В докладе обсуждаются возможные режимы использования комплекса КОЛЛИНГ и предлагается интерпретация полученных результатов.

ПОЛИКАРПОВ А.А., ТУЛДАВА Ю.А.

ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

I. Компьютерная лингвистика - это область теоретических и прикладных исследований и разработок, связанных с проблемами моделирования и исследования с помощью ЭВМ естественно-языковых систем коммуникации. Эта область порождается потребностями научно-технической революции, в частности потребностями построения систем искусственного интеллекта (ИИ). Проблема ИИ не может быть решена без учета лингвистических теорий, моделей и данных. Это, с одной стороны, связано с тем, что основной формой фиксации человеческих знаний является тот или иной естественный язык, точнее, его тексты. Эффективный поиск различной глубины и широты в море текстов с целью нахождения нужных сведений является обостряющейся проблемой современной цивилизации. С другой стороны, без лингвистических познаний не обойтись и при разработке систем общения человека с ЭВМ на естественном языке, что необходимо для повышения эффективности функционирования современных сложных человеко-машинных комплексов.

2. В свою очередь, с переходом на машинные методы сбора, хранения, поиска и классификации естественно-языковых данных лингвистика получает мощный импульс к дальнейшему развитию. Наступление эры ПЭВМ делает лавинообразным процесс компьютеризации лингвистики.

3. Собственные задачи лингвистики, которые могут решаться на ее компьютерной стадии, выражаются прежде всего в повышении эффективности процессов сбора, систематизации и классификации накопленного эмпирического знания в своей предметной области. Это выражается, прежде всего, в создании машинных фондов различных языков, включающих в себя в качестве главных компонентов текстовые и словарные БД. При этом обнаруживаются новые, часто удивительные факты. Во-вторых, задачи этого этапа развития лингвистики заключаются в новой постановке и проработке долгое время оставававшихся в тени фундаментальных вопросов, например, вопросов о механизме функционирования языка и сочленении языкового и неязыкового (энциклопедического) знания в актах человеческой коммуникации, а также о необходимости учета в общелингвистических моделях видов и характеристик памяти человека и вре-

менных параметров оперирования человека информацией. В целом, обнаруживается бесперспективность только формального моделирования языка в отрыве от функций и условий его существования. Вместо формально-структурного подхода сейчас вырисовывается постпозитивистский функциональный подход (Cluver, 1988). Еще более точно он может быть обозначен как постструктуралистский; системный подход в науке о языке (Мельников, 1978; Mel'nikov, 1988).

4. Высокоэффективным для проведения исследований и накопления новых данных и знаний являются экспертные системы. В лингвистике этот разряд систем еще не получил должного развития, за исключением области распознавания устной речи (см., например, систему Hearsay - II),хотя именно лингвистика как область, предмет которой обладает сложным строением и наличием множества вероятностных связей между ее единицами, является одним из наиболее подходящих приложений идеологии экспертных систем. Одной из проблем этого рода в лингвистике является, например, проблема атрибуции авторства текста. Приближением к этому классу систем становятся некоторые автоматизированные лексикографические системы (см., например, работы Л.И. Колодяжной), организующие труд лексикографа по сбору знаний из текстов и от экспертов,а также предоставляющие в распоряжение лексикологов и грамматистов удобное средство "перелопачивания" и статистических оценок материала словарей.Для этих систем вполне подходит определение "лексикографические процессоры". Только на основе лексикографических процессоров могут строиться многоаспектные, гибкие по организации, динамичные (открытые для пополнения - актуализации) автоматические словари.

5. Опыт эксплуатации первых поколений систем МП оказался чрезвычайно полезным для определения действительной значимости тех или иных лингвистических знаний. В частности, обнаруживается традиционная и перенесенная в новую область машинного перевода недооценка значимости такого языкового явления, как лексическая многозначность. Вследствие этого - ближайшая актуальнейшая потребность компьютерной лингвистики - оперативное построение и массовый выпуск в действие автоматических моно- и политематических семантических частотных словарей (словарей с частотами не только слов, но и их

значений, синонимических, гиперлексемных, родо-видовых и иных группировок слов), а также автоматических контекстуальных словарей (содержащих данные о значениях, актуальных для данной предметной области, а также о контекстуальных условиях их реализации и правилах применения этой информации для автоматического разрешения лексической многозначности).

6. Автоматические контекстуальные словари необходимы не только для МП, но и для любой другой области, имеющей дело с той или иной смысловой обработкой текстов - диалоговых систем, систем контент-анализа, развитых АИПС.

7. Информационный поиск в широких полнотекстовых базах данных на основе редуцированного лингвистического обеспечения и изощренные системы диалога человека с ЭВМ в той или иной узкой предметной области пока оказываются серьезно противопоставленными друг другу. Однако, с одной стороны, прогрессирующая интеллектуализация АИПС, превращающая информационный поиск в будущем в диалог пользователя с базой данных на естественном языке, а, с другой, - прогрессирующая универсализация диалоговых систем, расширение тематического охвата каждой из них, обещают сильное сближение этих двух типов систем. Собственно, некоторые современные лингвистические процессоры проектируются в равной степени пригодными для использования как в одной сфере, так и в другой.

8. Прогресс в области построения процессоров, предназначенных для МП,уже сейчас позволяет вести работы по созданию систем МП, действующих в реальном масштабе времени, например, МП устной речи или МП, осуществляющегося по ходу составления пользователем документа на родном языке.

9. Прогресс в области построения лингвистических процессоров позволяет сегодня поставить на практическую почву вопрос о разработке эффективного контроля орфографии, грамматики, лексико-семантических или лексико-грамматических свойств слов (см., например, работы С.А. Яблонского и его группы). На очереди интеллектуальные редакторы, позволяющие контролировать семантику и стилистику текста не хуже обычного редактора и даже - исправить допускаемые человеком ошибки.

РЕДЬКО Л.Ф.

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ИНТЕЛЛЕКТУАЛИЗАЦИЯ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

Применяемые в настоящее время средства и методы интеллектуализации вычислительных систем не в состоянии обеспечить эффективную обработку большого объема разнородных знаний. Причина неудачи кроется в предпринятой дорогостоящей попытке ограничиться сильно усеченным " естественным " языком. Между тем, именно отброшенные средства предназначены для обеспечения компактного представления знаний большого объема, манипулирования ими на любом уровне абстракции, конкретизации и обобщения, активного использования одних и тех же модулей, понятий для решения различных задач проблемной области.

В докладе рассматриваются основные преимущества средств, предоставляемых Е - языком для обработки и пополнения знаний естественным интеллектом и отсутствующих в системах искусственного интеллекта.

I. Сильно развитые средства именования и распознавания объектов, отношений, их свойств. В зависимости от необходимого уровня абстрагирования, конкретизации человек может присваивать одному и тому же объекту до семи имен. Отсутствие такой мощности средств в математике, языках программирования, языках описания данных приводит к обработке больших массивов данных как однородных, без учета особенностей подмножеств массивов, " индивидуальности " элементов либо к резкому сужению проблемной области.

2. Правила замены определений понятий на их имена, заданные в толковых словарях, инкапсулированные и недоступные для вмешательства, обеспечивают активность, модульность, компактность представления длинных выражений при конструировании новых.

3. Близость понятий, именованных на Е - языке, определяется в результате обращения к порождающим их типам. Понятия, порожденные одним семантическим типом содержат большое число подобных свойств.

Применение механизмов порождения Е - языковых конструк -

ций и их элементов, ограничений на порождаемые множества позволяет человеку отбрасывать недопустимые выражения, определять эквивалентность, близость анализируемых конструкций и их составляющих, динамически порождать имена объектов, отношений, их свойств, избегая на всех этапах комбинаторного взрыва.

Необходима громадная работа по " инвентаризации ", подробному описанию всех свойств Е – языка в соответствии с их прямым назначением : что , на каких этапах , как используется естественным интеллектом в процессе получения и обработки знаний. Е – языки отличаются степенью развитости того или иного отдельного средства. Сопоставление возможностей Е – языков различных групп должно привести к пониманию того, каким мог бы быть оптимальный язык, предназначенный для сильно интеллектуализированных компьютеров.

Практическим результатом исследований предложенной направленности в самое ближайшее время могут быть :

значительное повышение скорости и надежности автоматического распознавания устной и письменной / слитной / речи,
разработка требований к языкам программирования высокого уровня, языкам описания знаний и манипулирования ими,
разработка формальных средств интегрирования, пополнения знаний проблемных областей,
разработка эффективных средств принятия решений в экспертных системах.

Реструктурирование огромного объема знаний лингвистов с целью разработки эффективных механизмов представления и пополнения знаний в ЭВМ могло бы внести значительный , конструктивный ,вклад в решение одной из важнейших проблем искусственного интеллекта.

Л И Т Е Р А Т У Р А

I. Редько Л.Ф. Порождение конструкций естественного языка. Искусственный интеллект – 88. Переяславль – Залесский. 1988
2. Редько Л. Ф. Об инструментальной системе порождения адекватных конструкций естественных языков . " ЭВМ – перевод – 89 " Международный научно – технический семинар. Тбилиси. 1989.

СЕВАСТЬЯНОВ А.А., ФУНШТЕЙН С.Г., ШЕРШНЕВ В.Н. МОДЕЛЬ СМЫСЛООБРАЗОВАНИЯ В СИТУАЦИИ АФФЕКТИВНОГО ВОСПРИЯТИЯ ОБЪЕКТА

Разработчики систем управления нередко сталкиваются с фактами, когда эксперты (ЛПР) не могут дать мотивированной семантической интерпретации своего поведения.

По мнению авторов, аналогичным образом осуществлялось информационное взаимодействие субъекта с внешним миром в период знаковой доязыковой коммуникации. Дальнейшее развитие коммуникативного процесса обусловило утрату аффективными компонентами коммуникации приоритетной функции. Однако в ситуациях, когда стандартные языковые средства не позволяют выразить то, что ищет вербального выхода, в действие вступает "архаичный" механизм денотативно-аффективных классификаций, который переориентирует языковую систему, не давая ей "перегреться", а языковому сознанию отступить. Этот механизм, названный нами денотативно-аффективным классификатором (далее ДАК), разрешает для обозначения новых смыслов недопустимые с позиции семантической и синтаксической нормы случаи употребления знаков.

Авторы полагают, что в основе ДАК лежит выделенная В.Ф.Петренко способность аффекта выступать оператором перехода от предметно-денотативных к аффективно-коннотативным формам отражения, и за счет этого приводить к образованию новых семантических структур [I, с.78].

Очевидно, что механизмы типа ДАК могут использоваться в системах искусственного интеллекта, в частности в системах управления сложными объектами.

Л И Т Е Р А Т У Р А

I. Петренко В.Ф., Кучеренко В.В., Нистратов А.А. Влияние аффекта на семантическую организацию значения// Текст как психолингвистическая реальность. М., 1982 .

СИЛЬНИЦКИЙ Г.Г.

О СООТНОШЕНИИ КОЛИЧЕСТВЕННЫХ И КАЧЕСТВЕННЫХ КРИТЕРИЕВ КЛАССИФИКАЦИИ ЯЗЫКОВЫХ ЕДИНИЦ

I. Классификации могут строиться по экстенсиональным (денотативным) или интенсиональным (сигнификативным) критериям. В первом случае выявляются соотношения между объектами, во втором - между признаками объектов. Каждый признак (сочетание признаков) задает потенциальный класс объектов, характеризуемых наличием данного признака (сочетания признаков).

Сочетания признаков могут быть конъюнктивными и дизъюнктивными. Конъюнктивные сочетания ("комплексы") признаков усложняют, конкретизируют сигнификативное содержание класса и соответственно сужают его денотативный объем (количество охватываемых им объектов). Дизъюнктивные сочетания ("парадигмы") признаков, наоборот, обобщают сигнификативное содержание класса и соответственно расширяют его денотативный объем.

Любое сочетание (как конъюнктивное, так и дизъюнктивное) элементарных (принимаемых в данной классификации в качестве исходных) признаков может суммарно репрезентироваться одним "макропризнаком"; последние, в свою очередь, могут объединяться в макропризнаки более высокого порядка. Таким образом, классификация приобретает иерархическую структуру, на каждом уровне которой все множество рассматриваемых явлений распределяется по некоторому количеству классов, каждый из которых определяется соответствующим "признаковым основанием", т.е. некоторой совокупностью признаков, представленных у всех объектов данного класса.

2. Будем различать "гомонимический" (внутриуровневый) и "гетеронимический" (межуровневый) параметры классификации. В гомонимическом плане устанавливаются соотношения между классами (экстенсиональный критерий) или признаками (интенсиональный критерий) одного и того же уровня. В гетеронимическом плане осуществляется переход от одного иерархического уровня классов/признаков к другому, более абстрактному ("гиперонимический" подход) или более конкретному ("гипонимический" подход).

3. Основной тезис данного сообщения состоит в том, что чисто количественные методы классификации могут быть последовательно и однозначно применены лишь в г о м о н и м и ч е -

ском ее аспекте, тогда как любой гетеронимический подход, как правило, сопряжен с привлечением качественных, неквантифицируемых критериев. Используя философскую терминологию, можно сказать, что гетеронимический параметр представляет область перехода количественных классификационных критериев в качественные. Если рассматривать каждый классификационный уровень как относительно автономную подсистему, то данный тезис может трактоваться как частный случай теоремы о неполноте К.Геделя.

4. Указанный тезис может быть проиллюстрирован путем сопоставления эвристических возможностей методов корреляционного и кластерного анализа. Корреляционный анализ устанавливает статистически значимые соотношения между признаками, априорно принимаемыми за равноценные. Кластерный анализ группирует сами отношения между признаками, т.е. устанавливает отношения более высокого уровня обобщения (отношения между отношениями). Таким образом, корреляционный анализ может рассматриваться как гомонимическая (одномерная) процедура, кластерный анализ - как гетеронимическая (многомерная). Существенное методологическое различие между ними состоит в следующем.

Корреляционный анализ дает, при заданном математическом аппарате и определенном наборе объектов и признаков, однозначные результаты: относительно любой пары признаков может быть однозначно установлено, имеется ли между ними релевантная (положительная или отрицательная) связь. Кластерный анализ при тех же исходных условиях допускает столько различных группировок исследуемых явлений, сколько имеется иерархических ступеней в полученной дендрограмме; выбор какого-либо одного варианта определяется здесь критериями, не заложенными имманентно в самой применяемой процедуре, но привносимыми извне.

5. Корреляционный анализ имеет дело с конкретными качественными признаками изучаемых объектов, кластерный анализ - с абстрактными отношениями. Поэтому корреляционный анализ может быть определен как процедура формулирования исходных гипотез, интерпретирующих рассматриваемую эмпирическую область, кластерный анализ - как процедура проверки этих гипотез. Таким образом, кластерный анализ не должен следовать непосредственно за корреляционным, но должен основываться на предварительной качественной интерпретации корреляционных данных. Приведенные положения подкрепляются примерами.

СОКОЛОВА С.В.

ПРОГРАММАНАЯ РЕАЛИЗАЦИЯ ПРОЦЕДУР ПЕРЕВОДА В СИСТЕМЕ МАШИННОГО ПЕРЕВОДА MULTIS

MULTIS - многоязыковая система машинного перевода, функционирующая на персональных компьютерах типа PC XT, PC AT. Разработаны и внедряются англо-русская, франко-русская, испанско-русская и русско-английские подсистемы. Программное обеспечение MULTIS реализует функции ведения лингвистической базы данных системы и функции перевода. Лингвистическая база данных содержит двуязычные машинные грамматики и словари слов и словосочетаний. Процесс перевода поддерживается общими для всех подсистем процедурами, обеспечивающими ввод и вывод информации, диалог с пользователем и доступ к лингвистической базе данных.

Для организации собственно лингвистических процедур семантико-синтаксического перевода принят подход раздельного моделирования алгоритмов, обслуживающих уровень групп и уровень предложений.

Алгоритмы перевода групп - это алгоритмы, обеспечивающие анализ и перевод минимально связанных лексических единиц. Примеры выделяемых групп: сложные глагольные формы, простые именные группы, наречные группы, группы прилагательных, не входящие в состав простых именных групп и пр. Эти алгоритмы моделируются как восходящие алгоритмы разбора, основанные на атрибутных транслирующих LR(I) грамматиках и реализуются программно с помощью графовых структур, близких к расширенным сетям переходов.

На уровне предложений такой подход применить значительно сложнее, так как возникает колоссальная неоднозначность при применении правил разбора, кроме того, тексты на естественном языке как правило содержат опечатки, имена, названия, отсутствующие в словаре. Здесь применяется подход, основанный на использовании базовых структур - семантико-синтаксических фреймов, задающих типовые структуры предложений, элементами этих структур являются сегменты преложений с выделенными ядерными элементами. Сегменты могут представлять собой как группы, так и целые предложения. Работа с вложенными базовыми структурами поддерживается с помощью стекового механизма.

Разработан минимальный набор трансформаций базовых структур, описывающих изменение залога.

СОРОКОЛЕТОВ П.В.

СПЕКТРАЛЬНЫЕ МЕТОДЫ ДЛЯ РАСПОЗНАВАНИЯ И КЛАССИФИКАЦИИ СИМВОЛЬНЫХ ОБЪЕКТОВ

Одним из узких мест в системах обработки символьной информации является сопоставление структур. Алгоритмы символьного сопоставления служат основой для распознавания и классификации символьных объектов в программах обработки текстовых документов, искусственного интеллекта, морфологического и синтаксического анализа естественного языка, поддержки словарей и текстовых баз данных и др.

Повышение эффективности фундаментальных алгоритмов повлечет, таким образом, повышение продуктивности программирования в любой из указанных областей.

В рамках вычислительных систем (ВС) с традиционной Неймановской архитектурой и алгоритмов посимвольного сопоставления возможности качественного улучшения на сегодня исчерпаны. В качестве одного из альтернативных подходов можно предложить использование мощного аппарата спектрального анализа, созданного и развитого в области обработки сигналов [1, гл. 6,12] и физических наук [2, с.580-596].

Пусть $S_i = c_1\ c_2\ c_3\ \ldots\ c_m$ - произвольная строка, где c_j, $j = 1 \ldots m$ - символ, принадлежащий некоторому конечному алфавиту C_k :

$C_k = \{a_1, a_2, a_3, \ldots, a_k\}$, a_j, $j = 1 \ldots k$ - буквы алфавита.

Множество $S = \{S_1, S_2, \ldots, S_N\}$ есть множество подлежащих распознаванию строк. Поставим в соответствие алфавиту C_k числовое множество D_k так, чтобы каждой букве a_j соответствовало единственное кодовое число d_j. Добавим 0-й элемент: $\bar{D}_k = D_k \cup \{0\}$.

Заменив исходный символьный образ строки числовым, получим "дискретный сигнал" $D_i = \{d_1, d_2, \ldots, d_m\}$, который затем преобразуем следующим образом (используя аналогию с выборкой сигнала):

1. усиление по амплитуде;

$\bar{d}_j = 1/d_j$, если $d_j < z$

$\bar{d}_j = d_j * d_j$, если $d_j >= z$, где z - "порог усиления".

2. растяжение сигнала "по времени" путем вставки 0-го элемента: $\bar{D}_i = \{\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_m\} \rightarrow \tilde{D}_i = \{\bar{d}_1, [0,0, \ldots], \bar{d}_2, [0,0, \ldots], \ldots, \bar{d}_m\}$, где d - коэффициент масштабирования по времени, мени, есть число 0-х элементов, вставляемых между отсчетами сигнала.

Подвергнем сформированный D-образ дискретному преобразованию Фурье, переведя "временную" последовательность в "частотную":

$$X(к) = 1/m \sum_{j=1}^{m} \tilde{d}_j W^{jk}, \quad k = 1...m;$$

$$W = e^{-2\pi i/m}$$

Для каждого исходного S-образа получим его спектр, представленный ненулевыми гармониками $X_i(к)$, к = I...m; отбросив незначимые гармоники, меньшие порога усиления, еще уменьшим размерность исходной задачи. В результате для каждой строки будем иметь ряд из $L << m$ чисел, характеризующих S-образ. Сравнивая их, мы сможем судить о сходстве и степени сходства классифицируемых символьных объектов.

Отметим, что сходные методы были применены при сравнении изображений [3, стр.256-307], причем чем сложнее были анализируемые образы, тем надежнее происходило распознавание.

Практическая реализация метода будет целесообразной при двух условиях:

I. наличии арифметического сопроцессора в составе аппаратуры ВС, разгружающем центральный процессор;

2. диалоговом характере работы с текстовой информацией, позволяющем вычислять спектральные характеристики в темпе действий оператора.

Учитывая нечеткий характер распознавания, изначально налагаемый данным методом, его использование позволяет обрабатывать зашумленные тексты и символьные строки, например, фразы на естественном языке, содержащие орфографические ошибки. Метод можно использовать в сочетании с посимвольным сопоставлением в виде двухшаговой процедуры:

I. отбор кандидатов на основе спектрального анализа в большом пространстве текстов;

2. символьное сопоставление отобранных на шаге I кандидатов с образцом.

Л И Т Е Р А Т У Р А

I. Л. Рабинер, Б. Гоулд. Теория и применение цифровой обработки сигналов. - М.: Мир, I978. - 848 с.

2. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. - М.: Наука, I984. - 832 с.

3. Применение методов Фурье-оптики//Под ред. Г. Старка. - М.: Радио и связь, I988. - 420 с.

СУПРУН Л.С.

СТАНДАРТИЗАЦИЯ ВОПРОСИТЕЛЬНЫХ ПРЕДЛОЖЕНИЙ НА ВХОДЕ

I.При количественном анализе функционирования вопросительных предложений наблюдается неравномерное распределение (сплошная выборка из художественных и публицистических текстов).Максимум употреблений приходится на трех- и четырехэлементные предложения (вопросный оператор считается одним из элементов).При увеличении длины вопросительных предложений до I3 элементов, частотность уменьшается в 25 раз, следовательно в системе диалога предпочтительно пользоваться четырех- и трехэлементными предложениями.Вторая тенденция количественного распределения заключается в уникальности употребленного многоэлементного вопросительного предложения, так частотность трехэлементной модели КАК РО составляет 622 единицы, а у четырехэлементной модели КАК РОG этот показатель составляет 38, таким образом большему усложнению вопросительных предложений и связанному с ним увеличению разнообразия моделей после вопросных операторов соответствует меньшая встречаемость таких структур.

2.При организации вопросно-ответных систем наиболее рациональный путь – отказ от многоэлементных вопросительных структур и замена их семантически эквивалентными маломерными структурами на основе ядерных композиций, т.е. переход от систем большой мерности (многоэлементных структур, затрудняющих получение ответа) к системам малой мерности.

3.Следующим важным этапом в оптимизации вопросительных предложений на входе является расширение, "разворачивание" вправо и влево от вопросного оператора, т.к. не всегда вопросительное предложение начинается с вопросного оператора. Существенным моментом при создании вопросно-ответной системы является тот факт, что структура вопроса содержит большую часть структуры ответа в ситуации, когда вопрос задается пользователем к системе.Предметная область вытормозит в свою очередь элементы-расширители реализатора ответа.

4.Под стандартизацией вопросительных предложений мы понимаем приведение множества многоэлементных малоупотребительных, нечетких или омонимичных структур к определенным ядерным трех- и четырехэлементным композициям.Семантика вопросного оператора сузит круг элементов, следующих за вопросным оператором,определяя специфику составляющих компонентов.

ТАЭЛЬ К. Л.

ТЕКСТОВАЯ БАЗА ДАННЫХ ПРИ ИССЛЕДОВАНИИ ПОРЯДКА СЛОВ В ЭСТОНСКОМ ЯЗЫКЕ

Доклад посвящается опыту создания машинного фонда эстонских текстовых предложений, целью которого являлось статистическое исследование порядка слов в эстонском языке. Предметом исследования послужил самый нейтральный литературный язык, точнее, прозаические тексты на обыденные темы, которые были почерпнуты из прессы последних лет. Единицей анализа являлось простое предложение, которому в сложном предложении соответствует несамостоятельное предложение. Объем выборки составлял 3000 предложений. Достаточность такого объема выборки подвергалась проверке с помощью теста, проведенного финскими языковедами [3, с. 102 - 109]. Языковой материал был введен в ЭВМ в закодированном виде. Для кодировки был использован цифровой код, где каждая цифра обозначала определенную позицию в коде предложения, напр., позиция 17 обозначала структуру предложения. Такие позиции представляли собой переменные, которые каждый раз были заполнены цифрой, указывающей на конкретную лингвистическую категорию и обозначающей подкатегорию переменной. Использованы 74 позиции, с помощью которых предложение описывалось в целом с точки зрения его структуры, функций и типа, а отдельная фраза - с точки зрения ее синтаксических, семантических и инфоструктурных свойств [2].

В ходе работы была высчитана средняя частота встречаемости подкатегорий всех переменных и было описано распределение форм проявления переменных при выполнении определенных условий, т.е. в комбинации с другими переменными. Это позволило представить всесторонний статистический обзор порядка слов в текстовых предложениях эстонского языка.

Созданный машинный фонд является универсальным и может быть использован в разных целях многими исследователями, занимающимися проблемами синтаксиса и текста.

В данном докладе внимание сосредоточивается на попытке решения задачи взаимовлияния факторов, формирующих порядок слов. Была поставлена цель выяснить, какие признаки необходимы и ДОСТАТОЧНЫ для определения разных моделей порядка слов.

Мы опираемся на методику, разработанную А. Ивахненко,

т.е. на метод группового учета аргументов. Цель этой методики заключается в восстановлении модели оптимальной сложности [1, с.31 - 32], т.е. в селекции достаточного количества аргументов из большого количества вводных аргументов в целях получения желаемого вывода. Селекция происходит путем комбинирования аргументов по два, причем комбинации постепенно усложняются. Аргументами второго ряда становятся лучшие комбинации первого ряда и т.д.

Для решения поставленной задачи в качестве вводных аргументов послужили использованные переменные, а в качестве вывода - модель порядка слов. Целью являлось найти такие комбинации признаков, которые в возможно большей части были бы свойственны только одной модели порядка слов. В идеале это означало бы, что данная комбинация признаков стопроцентно определяет модель порядка слов.

Основным критерием селекции при проведенном анализе являлась свойственность признака для данной модели, которая выражается уравнением

$$\text{СООТНОШЕНИЕ} = \frac{X}{A} \qquad (1),$$

где A обозначает абсолютную частоту некого признака a во всем материале, а X - абсолютную частоту этого же признака в исследуемой модели, т.е. при данном выводе. Критерий СООТНОШЕНИЕ дополнен при анализе ограничением REL, котсрое показывает относительную частоту признаков в данном выводе и выражается уравнением

$$\text{REL} = \frac{X}{Z} \qquad (2),$$

где X обозначает, как и в уравнении (1), абсолютную частоту признака a в данной модели порядка слов и Z - абсолютную частоту данной моели порядка слов во всем материале. REL исключает выделение абсолютно периферийных комбинаций.

В докладе сосредоточивается внимание на практических проблемах данного анализа.

ЛИТЕРАТУРА

1. Ивахненко А., Зайченко Ю., Димитров В. Принятие решений на основе самооргаизации. М. 1976.

2. Таэль К. Автоматический анализ расположения элементов эстонских текстовых предложений. Диссертация на соискание ученой степени кандидата филологических наук. Таллинн 1989.

3. Hakulinen A., Karlsson F., Vilkuna M. Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus. Helsinki 1980.

ТИРМАСТЕ Ю.В., МИКК Я.А.

АНАЛИЗ СЛОЖНОСТИ УЧЕБНИКОВ АНГЛИЙСКОГО ЯЗЫКА НА ПЕРСОНАЛЬНОЙ ЭВМ

Желательно, чтобы сложность учебника возрастала постепенно, из года в год, а не скачкообразно. При составлении учебников возникает также проблема их словарного состава и т.п. Отсюда необходимость анализа уже существующих и составляемых учебников.

В Тартуском университете выработана система машинного анализа сложности учебников английского языка. В основу этой системы заложена программа М. Шулера. Исследования в этом направлении начались в США уже десятилетия назад. Однако там анализ проводился лишь по отдельным отрывкам учебника. Мы же для обеспечения большей достоверности решили проанализировать все его основные тексты.

Текст в ЭВМ можно ввести или с помощью сканнера (читающего устройства), или вручную. Мы использовали сканнер, однако затем вручную исправляли ошибки, возникшие из-за невысокого качества печатного текста, и отметили имена собственные. Затем с помощью специальной программы текст переводится в форму, необходимую для дальнейшего анализа. В результате ЭВМ выдает статистические данные о тексте (количество слов, предложений и т.д.), результаты анализа сложности по 10 формулам (методики Dale-Chall, Flesh-Kincaid, Coleman, Tuldava, Holmquist и др.), а также соответсвующую графическую информацию, используя для этого список из 3000 слов, наиболее часто встречающихся в лексике американских школьников. Программа выдает и слова, не встречающиеся в этом списке.

Специальная программа трансформирует анализируемый текст в форму для чтения в системе обработки базы данных (например, dBase), где можно составить частотный словарь анализируемого учебника, указав частотность каждого слова в каждом параграфе. С помощью такой системы мы установили, что скачек в сложности учебников от IУ к У классу для эстонских школ значительно больше, чем от У к УI классу. Мы распечатали также слова, встречающиеся в учебнике менее 6 раз, поскольку они могут оставаться неусвоенными многими школьниками. С помощью названной системы можно прогнозировать сложность учебника за 1-2 рабочие недели. Система работает на ПЭВМ IBM PC/XT/AT в Тартуском университете и внедряется в ГДР.

ТРОЙНОВА Т.Г.

ДИСТРИБУТИВНО-СТАТИСТИЧЕСКИЙ МЕТОД АНАЛИЗА ЛЕКСИКИ В РЕАЛИЗАЦИИ НА ЭВМ

I. Одним из путей автоматизации процесса анализа лексики в текстах является реализация на ЭВМ дистрибутивно-статистического метода (ДСМ). Настоящий доклад является продолжением серии исследований данной проблемы под руководством А.А. Поликарпова. Предшествующий вариант программы был разработан для реализации на ЭВМ серии ЕС (Ахмеджанов и др., 1990). В настоящее время ведется работа по реализации ДСМ на ПЭВМ PC/XT(AT).

2. Исходным этапом процедуры обработки текста является его перевод в файл базы данных. Текст размечается положительным и отрицательным словарем. При этом символьная строка, составляющая слово, может включать любые символы (цифры, знаки, латиница, кириллица), но должна содержать более 2-х символов. Из текста строится конкорданс. Словоупотребления, помеченные словами или словоформами "отрицательного списка" в него не попадают. В конкордансе приводится положительная помеченность, адресная и исходная статистическая информация по каждой словоформе.

3. На основе конкорданса строится словарь ключевых основ и гиперлексем, которыми в новом цикле могут быть помечены соответствующие словоупотребления, которые поступают в циклы получения дистрибутивных и статистических характеристик.

ТУЛДАВА Ю.А.

ОБ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ЛИНГВИСТИЧЕСКИХ ОБЪЕКТОВ

При автоматической классификации выделяются т.наз. естественные типы, т.е. группы объектов, наиболее сходные друг с другом по достаточно большому числу признаков. Особенностью классифицируемых лингвистических объектов является то, что их зачастую нельзя однозначно приписать к тому или иному классу. Классификация подобных объектов должна основываться на представлении о классах как о размытых множествах, к каждому из которых объект относится с некоторой степенью принадлежности. На практике это означает возможность построения классификационных систем, где классы могут частично покрываться. Наиболее известными в этой области являются особые методы кластер-анализа, реализуемые на ЭВМ, для классифицирования объектов к кластерам, разрешающим пересечение. В докладе сравниваются результаты классификации лингвистических объектов при применении двух различных алгоритмов кластер-анализа (реализованных в Вычислительном центре Тартуского университета). Первым рассматривается неиерархический алгоритм "ФОРЭЛЬ-2" [I]. Вторым рассматривается иерархический алгоритм $B_к$ [2] для осуществления к-кластеризации, т.е. для получения кластеров, которые могут покрываться в объеме до к - I объектов, принимая $I \leq к \leq n - 2$, где n - число исследуемых объектов. Отметим, что в последних разработках $B_к$-метод усовершенствован в том смысле, что вместо фиксированной степени покрываемости кластеры могут перекрываться различным числом элементов, причем степень пересекаемости определяется внутренней структурой данных [3].

ЛИТЕРАТУРА

I. Математическое обеспечение ЕС ЭВМ. Вып. IO. - Минск, 1976.

2. Р. Эзремаа. Общая теория конструирования кластер-системы и алгоритмы для нахождения их численных представлений // Труды Вычислительного центра. Вып. 42. - Тарту, 1978.

3. Р. Эзремаа. Конструирование послойной кластеризации // Труды Вычислительного центра. Вып. 53. - Тарту, 1986.

ФЕРЕНЦ-МОШИНСКАЯ М.М.

К ВОПРОСУ О ФРЕЙМОВОМ РАСПОЗНАВАНИИ ПАТЕНТНО-ТЕХНИЧЕСКОГО ТЕКСТА

Лексический анализ текста патентных описаний упрощается, если проводить его с помощью так называемых фреймов, используемых для автоматического реферирования и аннотирования патентов. Фрейм – это шаблон, функциональная карта, инструмент, позволяющий выбрать из словесной статьи необходимую для анализа и синтеза информацию в конкретном тексте. Накладывая фрейм на реальный текст, можно анализировать окружение лексической единицы в контексте.

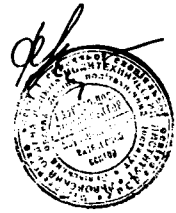
Используя такой инструмент, любой патентно-технический текст (патентное описание) можно представить в виде схемы-фрейма, состоящего из некоторого количества субфреймов, каждый из которых характеризуется определенным признаком или содержанием и реализуется при помощи определенных лингвистических средств, в первую очередь лексических.

Можно утверждать, что во все концептуальные поля и во все субфреймы патентных описаний вписывается некоторый стандартный набор лексических единиц, известный заранее, а поэтому легко поддающийся обработке как при переводе и реферировании, так и при автоматическом индексировании и аннотировании при наличии необходимого минимального словарного запаса, или при условии, что подготовлена соответствующая база данных в виде словарей, тезаурусов и грамматических правил.

Таким образом, функциональные карты или схемы способствуют детальному анализу патентно-технически текстов, узнаванию и конкретизации отрезков информации, а также синтезированию и обобщению. Предварительное знание лексического наполнения концептуальных полей этих фреймов позволяет установить семантические синтаксические валентности известных лексических единиц и их окружения.

Фреймовый подход к анализу смысла патентно-технического текста позволяет иметь достаточно высокую степень предсказуемости в использовании лексем по сравнению с собственно научным текстом, который обладает относительно высокой степенью предсказуемости по сравнению с художественным текстом.

В патентно-технических текстах подбор лексических единиц и их эквивалентов ограничен их композиционной структурой, которая находится в определенной зависимости от целей и задач каждого из подразделов этого текста.



ЦИЛОСАНИ Т.П.

О ФОРМАЛИЗМЕ МАКСИМАЛЬНОЙ ЭНТРОПИИ

Применение математических методов в лингвистике не всегда приносит желаемые результаты. Те или иные математические модели, являющиеся сами по себе "безупречными", в основном не оправдывают цель своего назначения. Поэтому, считаем необходимым обоснование их применения.

В работе дается попытка такого обоснования на примере выделения класса вероятностей, максимизирующих выражение шенноновской энтропии.

Наличие известной аналогии между количеством информации и термодинамической энтропией наводит на мысль о том, что в качестве математического аппарата при моделировании любых линейных структур может быть использован аппарат, аналогичный тому, который применяется в термодинамике (статистической физике).

Рассмотрение ряда вопросов статистической механики показывает "нефизичность", формальный характер основных положений статистической механики.

Пусть величина x принимает дискретные значения $x_i (i=1,2,\ldots,n;\ n \to \infty)$. Пусть нам не известны соответствующие вероятности P_i; все, что мы знаем - это математическое ожидание функции $f(x)$

$$\langle f(x) \rangle = \sum_{i=1}^{n} P_i f(x_i) \qquad (1)$$

На основе этой информации каково математическое ожидание функции $g(x)$?

С первого взгляда проблема кажется неразрешимой, т.к. имеющаяся информация недостаточна для определения вероятностей P_i. Уравнение (1) и условие нормировки

$$\sum_{i=1}^{n} P_i = 1 \qquad (2)$$

должны быть дополнены $(n-2)$ другими условиями, чтобы $\langle g(x) \rangle$ можно было точно определить.

Лапласовский принцип индиферентности является попыткой замены недостающего критерия выбора. Однако, за исключением случая, когда имеются налицо очевидные элементы симметрии, идея "равновозможности" событий может быть такой же произволь-

ной, как и любая другая. Поэтому, при оценках неизвестных средних величин этот путь был оставлен ввиду отсутствия конструктивного принципа, позволяющего выбирать некоторое распределение вероятностей по сравнению с другими, так же хорошо согласующимися с имеющейся информацией.

Хотя "объективистская" и "субъективистская" теории вероятностей математически идентичны, сами основные их концепции трудно объединимы. "Субъективистская" точка зрения более общая; соответствующая теория рассматривает и такие проблемы, которые с "объективистской" точки зрения лишены смысла. Проблема, сформулированная нами именно такого типа. Поэтому мы будем придерживаться "субъективистской" точки зрения.

Вернемся к проблеме, поставленной вначале и рассмотрим более общий случай. Пусть заданы средние значения функций $f_1(x), \ldots, f_m(x)$ где $m < n$. Задача заключается в оценке вероятностей $P_i = P(x_i)$, удовлетворяющих условиям:

$$\sum_{i=1}^{n} P_i f_k(x_i) = \langle f_k(x) \rangle = F_k \qquad (k = 1, 2, \ldots, m) \qquad (1')$$

$$\sum_{i=1}^{n} P_i = 1 \qquad P_i \geq 0 \qquad (2')$$

и максимизирующих шенноновское выражение для энтропии

$$S_I = H(P_1, \ldots, P_n) = -K \sum_{i=1}^{n} P_i \log P_i \qquad (3)$$

где K - произвольная постоянная. Эту величину мы называем энтронной распределения вероятностей P_i по аналогии с физической энтропией, т.к. их выражения формально идентичны.

Теперь очевидно, как решить нашу проблему; делая оценки на основе частичной информации, мы должны пользоваться распределением вероятностей, максимизирующим энтропию при условиях, соответствующих нашей частичной информации. Это единственный непредвзятый выбор распределения вероятностей, который мы можем сделать; использование любого другого распределения означало бы использование произвольного предположения, которое не должно делаться.

Все соотношения, полученные на основании предложенной модели внешне очень похожи на формулы статистической механики, однако ничего общего не имеют с физикой. Эти соотношения могут применяться всюду, где ситуация описывается уравнениями (1') и (2'). Т.о. они могут иметь применение в квантитативной лингвистике.

ЦИТКИНА Ф.А.

ПРЕДСТАВЛЕНИЕ РАЗНОЯЗЫЧНЫХ ТЕРМИНОЛОГИЙ В САМОАДАПТИРУЮЩЕМСЯ СЛОВАРЕ СИСТЕМЫ "АРМ ПЕРЕВОДЧИКА"

В докладе описываются цели, структура, объем, способы организации и подачи информации в первой версии системы автоматизированного рабочего места /АРМ/ переводчика, разработанной в Ужгородском госуниверситете на базе мини-ЭВМ СМ-4.

Под "АРМ переводчика" понимаем комплекс лингвистических и программных средств для облегчения труда переводчика; программное обеспечение АРМа в наши цели не входит. В отличие от других систем данного класса наш АРМ ориентирован на персональное, а не коллективное использование, он не предназначен для централизации информации, работает с меньшими ее объемами и требует иных конфигураций технических средств. Система состоит из лингвистических баз данных /ЛБД/, тематических баз данных /ТБД/, словаря, информационно-поисковой системы, пакета прикладных программ и системы управления базами данных. Наиболее активная часть системы - словарь. Он является терминологическим по наполнению, тематически-частотным по своей структуре и самоадаптирующимся по способу функционирования.

ЛБД в системе столько, сколько языков обслуживается. В файле ЛБД особую роль играет уникальный "номер" термина: он позволяет избежать дублирования информации и ускорить поиск в базах данных и словаре. Далее следует информация о наличие синонимов, указание на различные значения в случае полисемии, выделение ядерного слова и/или дериватов для поиска всего словарного гнезда.

Тематически-частотная организация словаря предполагает, что словарь разбит на тематические разделы, соответствующие тематике ТБД, а каждый раздел упорядочен по частотности терминов. Под частотностью мы здесь понимаем не традиционную частотность - частоту встречаемости в тексте, - а запросную частотность, т.е. частоту, с которой пользователь обращается к словарной статье.

Использование идеи уникального номера для связи соответствующих единиц разноязычных файлов словаря, а также конкретных необходимых пользователю ТБД и ЛБД позволяет легко наращивать число обслуживаемых языков и состыковывать АРМ с другими системами подобного типа.

ЦЫБ Е.А.

ЛЕКСИЧЕСКИЙ ОБРАЗ КОМПОЗИЦИИ ТЕКСТА (ОПЫТ АВТОМАТИЗАЦИИ ЛИНГВОПОЭТИЧЕСКИХ ИССЛЕДОВАНИЙ)

Композиция - это художественная форма в аспекте организованности, упорядоченности и в отвлечении от материала и эмоциональной окраски.

Предмет сообщения - как композиция поэмы А.Твардовского "По праву памяти" отражается в распределении лексики по ее компонентам. С помощью ЭВМ были получены словоуказатели и количественные характеристики распределения разных слов и двух важнейших их категорий, существительных и глаголов, по структурным частям поэмы, см. табл. 1.

Структрные компоненты				Слова и их категории		
Введ.	Гл.1	Гл.2	Гл.3	Всего	Сущ.	Глаг.
+	-	-	-	21	8	7
-	+	-	-	158	57	46
-	-	+	-	380	156	102
-	-	-	+	230	71	70
-	+	+	-	32	15	1
-	-	+	+	65	33	5
-	+	-	+	13	5	1
-	+	+	+	57	11	5
+	+	+	+	23	-	2
Всего				979	356	241

Композиционный анализ художественного произведения должен установить, на какие части расчленяется произведение, какую роль играют в его членении отдельные факторы материала, каковы способы разграничения частей и их объединения в художественное целое .

При обсуждении результатов, описанных в таблице, мы будем следовать следующей тактике. Вначале будем рассматривать характеристики отдельности структруных частей поэмы (противопоставленность каждой части всем другим). После этого будем анализировать выражение связей между ними. Затем рассмотрим вклад существительных и глаголов в композицию данного произведения.

Отдельность глав. Из таблицы видно, что пик лексического богатства поэмы приходится на вторую главу: своеобразие лексики в восемь раз увеличивается от Введения к первой главе, во второй главе оно в два раза больше, чем в первой, в третьей главе лексическое своеобразие уже слабее в 1,6 раз по сравнению со второй главой.

Связь глав. Можно считать, что связанность глав в первом приближении пропорциональна количеству общих слов. С этой точки зрения связь всех четырех частей оказывается слабее, чем для любых двух соседних глав, равно как и для трех последних компонентов. Теснее всего связаны вторая и третья главы. Общих слов у них в два раза больше, чем в смежных первой и второй и в пять раз больше, чем в несмежных первой и третьей главах. Введение резко противопоставлено всем трем частям. Нет таких полнозначных слов, которые объединяли бы Введение с другими главами поэмы.

Перейдем к описанию того, как композиция поэмы отражается в распределении существительных по структурным компонентам. Распределение существительных еще более контрастно демонстрирует противопоставленность Введения всем остальным главам. Нет ни одного существительного связывающего Введение с ними. Теснее связаны вторая и третья главы, и эта связь показана еще более выразительна. Теперь общих слов у них в три раза больше, чем в первой и второй главах и в шесть раз больше, чем в несмежных первой и третьей главах, и в три раза больше, чем во всех трех частях. Лексическое своеобразие по-прежнему возрастает от Введения к первой главе, затем ко второй, после чего происходит его спад.

Если обратиться к глаголам, то нетрудно убедиться, что композиция поэмы отражается в их распределении существенно слабее. Действительно, Введение опять противопоставлено всем другим частям, но не так ярко, как в слнчае с существительными. Связанность второй и третьей глав как бы воспроизводит связанность всех трех глав. Таким образом, распределение глаголов показывает скорее непрерывность текста, чем его расчлененность.

Качественные характеристики лексической демонстрации композиционной структуры изучаемого поэтического текста предмет особого рассмотрения.

ЧЕРКАСОВА Г.А.

КОМПЬЮТЕРНЫЙ АССОЦИАТИВНЫЙ СЛОВАРЬ РУССКОГО ЯЗЫКА

Институтом русского языка АН СССР совместно с группой психолингвистики Института языкознания АН СССР проводится ассоциативный эксперимент, целью которого является создание ассоциативного тезауруса русского языка.

В процессе опроса испытуемые указывают свой пол, возраст и специальность, записывают свою первую ассоциацию (реакцию) на все стимулы, содержащиеся в анкете.

Ассоциативный эксперимент проводится в несколько этапов, технология обработки которых включает следующие шаги:

1. Формирование исходного списка стимулов в памяти ЭВМ.

2. Программная генерация и печать анкет, содержащих по сто случайно выбранных из исходного списка стимулов.

3. Ввод заполненных в результате опроса анкет в память ЭВМ, проверка правильности набивки, исправление ошибок и формирование массива исходных файлов.

4. Сортировка и слияние данных исходных файлов, создание базы ассоциативного тезауруса.

5. Исследование накопленных в базе данных, формирование выборок и получение производных баз.

6. Анализ полученных в результате исследований данных и подготовка материалов для следующего этапа.

В процессе функционирования этой технологии создаются компьютерный ассоциативный тезаурус и словарь русского языка. Последний представляет собой сгруппированное и упорядоченное множество пар "стимул-реакция" и соответствующие им параметры респондентов.

Ассоциативный словарь существует только в компьютере на магнитных носителях и входит в программно-источниковый пакет "Лексико-семантических ассоциаций носителей русского языка". Его основное отличие от книжной формы в том, что он не является застывшим образованием, а постоянно изменяется и пополняется. При этом некоторые этапы его развития, которые могут представлять интерес для исследователей, фиксируются в виде файлов на магнитных носителях и могут быть распечатаны или подготовлены для издания в виде книги.

Первоначально результаты ассоциативного эксперимента обрабатывались в среде ЭВМ СМ-4. Разработана файловая система и пакет прикладных программ обслуживающих ее. Сформировано

28 исходных файлов, которые объединяют анкеты по специальностям опрашиваемых. Комплекс программ позволяет вводить анкеты, упорядочивать и добавлять их в словарь, организовывать накопленные данные в новые структуры и получать в виде распечаток любые фрагменты и результаты.

В настоящее время создается версия компьютерного ассоциативного словаря на персональном компьютере класса IBM-PC с использованием СУБД FOXBASE. Разработана и реализована технология переноса данных из СМ ЭВМ в среду ПЭВМ и обратно.

На первом этапе ассоциативного эксперимента использовалось 1285 стимулов, включающих около 900 лексем, некоторые из которых даны в разных формах и с предлогом "о". В базу введено 1200 анкет, т.е. на каждый стимул имеется почти 100 реакций. После сортировки и слияния исходных файлов получены и распечаны данные словаря в двух формах: "прямой" и "обратный" словари. Сформированы также различные выборки.

"Прямой словарь" образует множество стимулов, включенных в анкеты. Его словарная статья содержит все реакции, полученные на этот стимул, в порядке убывания частоты их встречаемости.

"Обратный словарь" - это список всех разных реакций, полученных в ассоциативном эксперименте, в алфавитном порядке. Он включает почти 26 тысяч словарных статей, каждая из которых содержит стимулы, породившие эту реакцию, в порядке убывания частоты пары "стимул-реакция".

При формирования выборок использовались следующие критерии: частота пары "стимул-реакция" в базе; количество стимулов, породивших реакцию; частота встречаемости реакции и другие.

Перспектива словаря предполагает развитие вширь за счет добавления новых стимулов и вглубь за счет получения большего количества реакций на имеющиеся стимулы. Планируется провести несколько этапов, на каждом последующем будет расширяться список стимулов за счет слов, содержащихся в реакциях.

Часть исходного списка стимулов (около 800 слов) использовалось для генерации новой серии анкет с тем, чтобы получить на каждый стимул 500 реакций. В настоящее время проводится обработка этих результатов.

В докладе рассматриваются результаты создания и анализа базы компьютерного ассоциативного словаря русского языка.

ЧЕРКАСОВА Г.А., ДМИТРИЕВА О.Л.

КОМПЬЮТЕРНАЯ ТЕХНОЛОГИЯ ЛЕКСИКОГРАФИРОВАНИЯ ОРФОГРАФИИ

В Машинном фонде Института русского языка АН СССР орфографическая словарная база данных создана как новый автоматизированный словарь /далее - ОСА/. Единица хранения ОСА - вокабула, обеспечивающая адекватную орфографическую информацию. На машинных носителях реально находятся 110 тысяч вокабул. Начальный информационный массив ОСА занесен на машинные носители по тексту 26 стереотипного издания Орфографического словаря русского языка. ОСА является базовой частью словарно-ориентированного орфографического подфонда Машинного фонда русского языка, предназначенного для хранения, исследования и использования автоматизированного орфографического словаря в различных прикладных и исследовательских аспектах. Орфографический подфонд, кумулятивный по существу, открыт для постоянного уточнения и новой информации.

Потребность в создании ОСА обусловлена, прежде всего, установкой на новую перспективную издательскую технологию - автоматизацию редакционно-издательской деятельности. Словарные работы автоматизированы на всех этапах - создание, редактирование и издание рукописи. В системе ОСА автоматизированы и находятся в стадии автоматизации следующие лексикографические процедуры формирования словника - накопление первичных данных в словарной базе, регулярное уточнение словника в целях соблюдения орфографического режима русского языка, развития информационного орфографического обеспечения. Осуществляется также синтаксический и семантический контроль содержания словаря - его структуры и единиц хранения в словарной базе.

Реальная эксплуатация системы может быть шире, чем усовершенствования издания. Компьютерный доступ к большому массиву орфографической информации предполагает и автоматизированную таксономическую проекцию словаря, классификацию по заданным параметрам, критериям. В тексте словаря можно выделить ряд признаков и автоматически осуществить выбор на следующих уровнях: морфологическом, синтаксическом, семантическом, лексическом, системном, историко-хронологическом, орфоэпическом, стилистическом, нормативном. Параметризация в ОСА позволяет кроме выбора из словарной базы по заданным значениям признаков также формирование списков однородных в каком-либо отношении лексем и статистические исследования русской орфографии.

ЧИАЛАШВИЛИ Е.Ш.

О СООТНОШЕНИИ ИНФОРМАЦИОННОЙ НАСЫЩЕННОСТИ И ИНФОРМАТИВНОСТИ ТЕКСТА ПРИ ПЕРЕВОДЕ

I При сопоставлении информационной насыщенности и информативности текста последний толкуется прежде всего как информация о концепции. Разнопорядковые текстовые структуры в разной мере значимы с точки зрения ее передачи реципиенту.

II Подсчет информационной насыщенности текста как величины абсолютной, подразумевающей общее количество информации в тексте, зависящее от длины текста и разнообразия его словаря, возможен как путем применения вероятностно-статистических методов, так и исходя из теории актуального членения.

III Информативность текста как величина относительная подразумевает потенциальную интерпретационную характеристику текста, позволяющую прогнозировать степень адекватности интерпретации смыслового восприятия текста реципиентом.

IУ Интерпретация может считаться адекватной лишь при условии, что реципиент толкует основную цель сообщения в соответствии с замыслом коммуникатора. Иными словами, если есть информация ($\mathcal{I}$), коммуникатор ($\mathcal{C}$), который воспринимает (P_c) информацию ($\mathcal{I}$), а затем с помощью языковых средств описывает ($\mathcal{R}$) ее, если это описание ($\mathcal{R}$) затем воспринимает (P_r) реципиент (r), то восприятие реципиента должно быть равно восприятию и описанию коммуникатора: $P_c = \mathcal{R} = P_r$.

У При переводе текста с одного языка на другой процесс усложняется. Необходимо соответствие следующего характера: $P_c = \mathcal{R} = P_r = Pt = P_r t = P$, где Pt – восприятие переводчика, Rt – его описание, а $P_r t$ – восприятие реципиентом текста-перевода.

УI Интерпретационная деятельность реципиента рассматривается, с одной стороны, как внешняя деятельность (в виде чтения или слушания), с другой стороны, как внутренняя (в виде мыслительной деятельности). Адекватность интерпретационной деятельности реципиента цели, поставленной коммуникатором, может зависеть от степени приобщенности реципиента как к денотату сообщения, так и к языку, используемому для его описания.

УII Существенным фактором интерпретационной деятельности является также и влияние конкретного языка на процесс моде-

лирования денотата. В соответствии с принципом лингвистической дополнительности /I, с 50/ мы исходим из того, что при моделировании объективной действительности в сознании человека переплетаются две картины этой действительности – концептуальная и языковая. Если концептуальная модель у носителей разных языков является инвариантной, то языковые модели отличаются друг от друга. Основную информацию реципиент воспринимает при помощи концептуальной модели. Языковая же модель вносит дополнительную информацию, что особенно наглядно проявляется при семантическом сопоставлении текстов разносистемных языков.

УШ.В качестве конституентов информативности рассматриваются такие экстралингвистические параметры как: гипотетический коэффициент информативности текста (степень словесной избыточности текста), а также предикативная структура текста (типы ее развертывания).

IX Таким образом, текст рассматривается как система элементов, функционально объединенных общей концепцией в единую структуру. В качестве конституирующих признаков текста выступают его целенаправленность, качественное изменение той информационной нагрузки, которую несет вне данной структуры каждый из ее элементов, а также иерархичность предикативной структуры (линейная структура, структура с постоянной темой, деривационная структура).

X Семантический анализ предполагает следующую типологию текстов: денотативные, десигнативные, денотативно-десигнативные, денотативно-оценочные, денотативно-предписывающие, десигнативно-оценочные, десигнативно-предписывающие; при прагматическом же анализе используется классификация текстов на фактологические, абстрактные и суперабстрактные.

Л И Т Е Р А Т У Р А

I. Брутян Г.А. Принцип лингвистической дополнительности "Философ. науки", I969 № 3.
2. Жинкин Н.И. Механизмы речи. М., I958.
3. Смысловое восприятие речевого сообщения, под.ред. Дридзе Т.М. и Леонтьева А.А. М., I976.
4. Чиалашвили Е.Ш. О роли актуального членения в смысловом восприятии русского и грузинского текстов. Труды ТГУ Тб., I98I.

НЙМ Х.Я.

ВОЗМОЖНОСТИ МОДЕЛИРОВАНИЯ МЕТОДОВ КОНВЕРСАЦИОННОГО АНАЛИЗА НА ЭВМ

При построении машинных моделей естественно-языкового общения в искусственном интеллекте или компьютерной лингвистике за теоретическую основу обычно берутся идеи теории речевых актов. Структура речевого акта как единицы общения в таких моделях типично представляется в виде фрейма, где фиксированы предусловия, /коммуникативная/ цель и последствия акта.

Подход теории речевых актов к моделям естественно-языкового общения, как известно, остро критикуется со стороны конверсационного анализа. Основная причина критики - чрезмерная абстрактность речевых актов как единиц общения и неучитывание роли контекста в интерпретации реального процесса общения. В конверсационном анализе разработаны свои принципы членения текста в единицы и характеризации этих единиц с точки зрения их функций.

Однако речевые акты как концептуальные единицы, с одной стороны, и единицы конверсационного анализа, с другой, можно также рассматривать как представляющие два уровня анализа общения. Уровень речевых актов представляет собой базисный уровень, где задаются базисные типы единиц общения вместе с принципами, регулирующими их использование. Уровень конверсационного анализа - это уровень реализации, где учитываются конкретные контекстуальные условия и фиксируются языковые средства реализации речевого акта в данных условиях.

В докладе рассматривается возможности реализации такой модели в виде компьютерной системы.

ЯБЛОНСКИЙ С. А.

ПРОЦЕССОР РУССКОГО ЯЗЫКА РУССИКОН-I

Необходимость создания процессора русского языка для персональных компьютеров отмечается рядом исследователей. За рубежом в области создания процессоров романских и германских языков получены существенные результаты. Достаточно отметить такие промышленные системы, как WordPerfect, Grammatik III, IV и др.

В настоящей работе приводятся результаты создания процессора русского языка РУССИКОН-I для персональных компьютеров типа IBM PC.

Процессор состоит из следующих основных блоков:

1. Система ведения, обработки и компрессии словарей включает комплекс машинных словарей: базового словаря русского языка (два варианта реализации: словарь Засориной - 40 тыс. словоизменительных основ и словарь Зализняка - 100 тыс. словоизменительных основ), словарей флективных классов (ФК), словообразовательных классов (СК), суффиксов и сочетаний суффиксов, псевдосуффиксов, префиксов и т. д. Разработана система обработки "новых слов", позволяющая в диалоговом режиме определить номер ФК, длину словоизменительной и словообразовательной основ слова, префикс, номер СК и ряд других характеристик. Базовый словарь русского языка реализован в виде B-дерева и компрессированного последовательного файла на магнитном диске.

2. Морфологический анализатор реализует алгоритм морфологического анализа, учитывающий явления словоизменения, словообразования и чередования гласных и согласных в основах слов. Отличительными чертами реализации алгоритма являются:

- учет омонимии на уровне выделения всех словообразовательных основ словаря, совпадающих со словообразовательной основой анализируемого слова;
- использование алгоритма приближенного морфологического анализа слов, для которых не найден словарный эквивалент основы;

- относительно высокий процент правильного разбора слов (90%), словообразовательные основы которых представлены в базовом словаре за счет использования словообразовательных классов при отождествлении слова со словарной основой и последующего предсинтаксического анализа предложения;

- скорость морфологического анализа - 10-50 слов/сек в зависимости от модели IBM PC и ее технических характеристик.

3. Синтаксический анализатор осуществляет анализ предложений в следующей последовательности:

- предсинтаксический анализ, состоящий из проверки более 200 правил предсинтаксиса, включающих проверку согласования слов в предложении, модели управления предлогов и глаголов;

- представление сложных предложений по возможности в виде композиции простых предложений на основе метода [1] и построения для простых предложений деревьев зависимости, используя подход [2];

- выделение именных словосочетаний.

4. Корректор русского языка, позволяющий выявлять ошибки правописания и согласования слов в предложениях проверяемого текста.

Процессор РУССИКОН-I используется при создании систем автоиндексации, для осуществления контроля ввода информации в базы данных и в текстовых редакторах, в системах распознавания русского текста, в сканерах, в качестве программной модели аппаратной реализации процессора русского языка в виде СБИС и др.

Процессор реализован на языке СИ в среде MS-DOS.

Л И Т Е Р А Т У Р А

1. Итоги науки и техники / сер. Информатика, т.8 - М.: ВИНИТИ, 1984.

2. Сухотин Б.В. Оптимизационные методы исследования языка. - М.: Наука, 1976.

УДК 681.3I

ЯЩЕНКО Н.Ю.

РАЗРАБОТКА РОЛЕВОЙ СЕМАНТИЧЕСКОЙ ГРАММАТИКИ ДЛЯ СИСТЕМ ПРЕОБРАЗОВАНИЯ ЕСТЕСТВЕННО-ЯЗЫКОВОЙ ИНФОРМАЦИИ В ОБРАЗНУЮ

Обсуждается связь логических аспектов представления знаний в системах преобразования текстовой информации в образную с построением требуемой семантики.

Понимание описания сцен означает способность преобразовать символьно-вербальную информацию в визуальное аналоговое представление. Адекватность аналогового представления статических сцен в решающей степени зависит от построения отношений на уровне локального контекста. Аналоговое представление динамических сцен связано с анализом семантики предикатных конструкций. Таким образом, целесообразно выделить концептуальный, эпистемологической и логический уровни в многослойной архитектуре лингвистического процессора, где слои определяются как отображения на уровнях абстракции данных, с использованием знаний, локализованных в соответствующем слое программного обеспечения.

Лингвистический процессор осуществляет построение внутреннего фреймового семантического представления (FR) с учетом прагматических целей системы на основе разработанной ролевой семантической грамматики: $GR: T \to FR$, где $T = \{t_1, \ldots, t_n\}$ - входная цепочка символов, состоящая из слов ЕЯ.

Базисными типами являются: σ - существительное в широком смысле слова; τ - предложение. Формально GR есть построение вычислительной модели для следующих отображений: $\varepsilon_{\sigma\sigma}: V_{\sigma\sigma} \times V_\sigma \to V_\sigma$; $\varepsilon_{\sigma\tau}: V_{\sigma\tau} \times V_\sigma \to V_\tau$; $\varepsilon_{\tau\sigma}: V_{\tau\sigma} \times V_\tau \to V_\sigma$; $\varepsilon_{\tau\tau}: V_{\tau\tau} \times V_\tau \to V_\tau$

которые интерпретируются соответственно: $\varepsilon_{\sigma\sigma}$ - преобразователь существительного в существительное в широком смысле слова; $\varepsilon_{\sigma\tau}$ - преобразователь существительного в предложение; $\varepsilon_{\tau\sigma}$ - преобразователь предложения в существительное и $\varepsilon_{\tau\tau}$ - преобразователь предложения в предложение.

В качестве примера определим синтаксическое выражение для отображения $\varepsilon_{\sigma\tau}$. Предикат P в общем случае может иметь n аргументных мест.

$Val(P(t_1,\ldots,t_n)) \rightleftharpoons (\varepsilon(\bar{P}, (Val(t_1), \ldots, Val(t_n))))^{Asg}$, где

$\bar{P}$ - интенсионал, а $\varepsilon(x_1, \ldots, x_{n+1}) = \varepsilon(\varepsilon(x_1, x_2), \ldots, x_{n+1})$

для соответствующих $x_1, \ldots, x_{n+1}$ и каждое вхождение ε означает соответствующее $\varepsilon_{\sigma\tau}$.

Различаем около 30 падежных позиций, для каждой из которых предусмотренной в модели предиката, даются семантические сведения о существительных, которые способны выступать в данной позиции, то есть, $\bar{S}_i : (\mu_{i1}, \ldots, \mu_{ie})$, где μ_{ip} - тип семантической характеристики для i -ой падежной позиции. Существительные же имеют соответствующие семантические коды в своих словарных статьях. Функция, характеризующая объект $t_j : \sigma$ в момент $\kappa \in Asg$ есть отображение

$\bar{\varphi}_{t_j}(\kappa) : |\delta_1, \kappa| \times \ldots \times |\delta_n, \kappa| \times |\eta_1, \kappa| \times \ldots \times |\eta_e, \kappa| \to B$, где

δ_m - тип морфологической характеристики объекта t_j ,

η_p - тип семантической характеристики объекта t_j ,

$B = \{true, false\}$.

Рассмотрим предикаты $\bar{\theta}(\bar{q}_{i1}, \bar{z}_{j1}), \ldots, \bar{\theta}(\bar{q}_{ie}, \bar{z}_{je})$, для которых $\nu = \overline{1,e}$, а $\bar{q}_{i\nu} \in |p_\nu, k|$, $\bar{z}_{j\nu} \in |\eta_\nu, \kappa|$ - значение семантических характеристик, имеющих одинаковые или подобные типы.

Каждое i -ое вхождение $\varepsilon_{\sigma\tau}$ соответствует приписыванию объекту t_j ролевой метки ρ_i.

$\varepsilon^i_{\sigma\tau} : |\eta'_1, \kappa| \times \ldots \times |\eta'_\ell, \kappa| \to B$, где $|\eta'_j, \kappa|$ - производный домен j -го семантического признака.

$$\eta'_{j\nu} = \inf_{\bar{z}_\nu \in H_{\eta_\nu}} (\bar{\varphi}_t(K)(\bar{z}_{1\nu}, \ldots, \bar{z}_{j-1\nu}, \bar{z}_{j+1\nu}, \ldots, \bar{z}_{\ell\nu})\, \bar{\theta}\, \bar{q}_{i\nu})$$

Таким образом, $Val(P(t_1, \ldots, t_n), \kappa) = \varepsilon(\bar{P}(\kappa), \bar{t}_1(\kappa), \ldots, \bar{t}_n(\kappa))$, то есть вычисление экстенсионала "целого" сводится к вычислению экстенсионалов "частей".

Аналогично построены синтаксические выражения для отображений $\varepsilon_{\sigma\sigma}$, $\varepsilon_{\tau\sigma}$, $\varepsilon_{\tau\tau}$.

На основе разработанной ролевой семантической грамматики разработан синтактико-семантический анализатор фраз ЕЯ-текстов.

АКТУАЛЬНЫЕ ПРОБЛЕМЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ.
Тезисы докладов Всесоюзной конференции.
Тарту 29-31 мая 1990 г.
На русском языке.
Тартуский университет.
ЭССР, 202400, г.Тарту, ул.Юликооли, 18.
Ответственный редактор К. Лепа.
Подписано к печати 12.04.1990.
Формат 60x84/16.
Бумага ротаторная.
Машинопись. Ротапринт.
Условно-печатных листов 9,07.
Учетно-издательских листов 8,76. Печатных листов 9,75.
Тираж 295.
Заказ № 252.
Цена 1 руб. 80 коп.
Типография ТУ, ЭССР, 202400, г.Тарту, ул.Тийги, 78.